TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY

Antti Hurmalainen

Robust Speech Recognition with Spectrogram Factorisation

Julkaisu 1249 • Publication 1249

Tampere 2014

Tampereen teknillinen yliopisto. Julkaisu 1249
Tampere University of Technology. Publication 1249

Antti Hurmalainen

Robust Speech Recognition with Spectrogram Factorisation

Thesis for the degree of Doctor of Science in Technology to be presented with due permission for public examination and criticism in Tietotalo Building, Auditorium TB109, at Tampere University of Technology, on the 9th of October 2014, at 12 noon.

Tampereen teknillinen yliopisto - Tampere University of Technology
Tampere 2014

Supervisors:

Dr. Anssi Klapuri,
Department of Signal Processing,
Faculty of Computing and Electrical Engineering,
Tampere University of Technology,
Tampere, Finland.

Dr. Tuomas Virtanen (Custos),
Department of Signal Processing,
Faculty of Computing and Electrical Engineering,
Tampere University of Technology,
Tampere, Finland.

Pre-examiner:

Dr. Shinji Watanabe,
Mitsubishi Electric Research Laboratories,
Cambridge, MA, USA.

Pre-examiner and opponent:

Dr. Emmanuel Vincent,
INRIA Nancy – Grand Est,
Villers-lès-Nancy, France.

Opponent:

Dr. Ville Hautamäki,
School of Computing,
Faculty of Science and Forestry,
University of Eastern Finland,
Joensuu, Finland.

ISBN 978-952-15-3371-6 (printed)
ISBN 978-952-15-3386-0 (PDF)
ISSN 1459-2045

Abstract

Communication by speech is intrinsic for humans. Since the breakthrough of mobile devices and wireless communication, digital transmission of speech has become ubiquitous. Similarly distribution and storage of audio and video data has increased rapidly. However, despite being technically capable to record and process audio signals, only a fraction of digital systems and services are actually able to work with spoken input, that is, to operate on the lexical content of speech. One persistent obstacle for practical deployment of automatic speech recognition systems is inadequate robustness against noise and other interferences, which regularly corrupt signals recorded in real-world environments.

Speech and diverse noises are both complex signals, which are not trivially separable. Despite decades of research and a multitude of different approaches, the problem has not been solved to a sufficient extent. Especially the mathematically ill-posed problem of separating multiple sources from a single-channel input requires advanced models and algorithms to be solvable. One promising path is using a composite model of long-context atoms to represent a mixture of non-stationary sources based on their spectro-temporal behaviour. Algorithms derived from the family of non-negative matrix factorisations have been applied to such problems to separate and recognise individual sources like speech.

This thesis describes a set of tools developed for non-negative modelling of audio spectrograms, especially involving speech and real-world noise sources. An overview is provided to the complete framework starting from model and feature definitions, advancing to factorisation algorithms, and finally describing different routes for separation, enhancement, and recognition tasks. Current issues and their potential solutions are discussed both theoretically and from a practical point of view. The included publications describe factorisation-based recognition systems, which have been evaluated on publicly available speech corpora in order to determine the efficiency of various separation and recognition algorithms. Several variants and system combinations that have been proposed in literature are also discussed. The work covers a broad span of factorisation-based system components, which together aim at providing a practically viable solution to robust processing and recognition of speech in everyday situations.

Preface

This work has been carried out at the Department of Signal Processing, Tampere University of Technology, during 2010–2014. Apart from the author, many people have contributed to the research and supporting work that eventually resulted in completion of this thesis.

First, I want to thank Anssi Klapuri, who actively encouraged me to pursue these studies and soon became my first supervisor in 2009–2010. This role was later transferred to Tuomas Virtanen, who has provided invaluable expertise and insight throughout these studies, and had an integral role in refining the publications to their final form. The results as presented in this thesis were reviewed by the pre-examiners, Emmanuel Vincent from INRIA and Shinji Watanabe from MERL. Emmanuel and Ville Hautamäki from University of Eastern Finland also agreed to act as opponents in the public defence.

I have repeatedly received assistance from colleagues at TUT, Radboud Universiteit Nijmegen, KU Leuven, TU München, and other institutions. Already the number of co-authors of this research is too large to be repeated here. Nevertheless, a special mention belongs to Jort Gemmeke, whose earlier, joint and simultaneous work is closely related to practically every aspect of this thesis. Many novel ideas, triumphs and tribulations were shared in countless e-mails all the way into the small hours before looming deadlines.

The practical side of the life of a researcher was secured by TISE graduate programme, our highly competent secretaries, knowledgeable IT support, and other personnel at the department and TUT. The audio research group, including but not limited to Pasi, Joonas, Hanna, Elina, Jouni, Marko, Konsta, Mikko, Toni(s), Annamaria, Katariina, Alexandr, Tom and numerous visiting researchers was always around to share the joys, woes, and collective knowledge of audio research.

Finally, I want to express my gratitude to people including Sami, Jeffrey, Niklas, Osbourne, James, Matt and Simon, who taught me the essence of audio signal processing long before it turned into an academic career.

Antti Hurmalainen
Tampere, 2014

Contents

List of Figures

List of Abbreviations

ANN	Artificial neural network
ASR	Automatic speech recognition
CNMF	Convolutive non-negative matrix factorisation
DFT	Discrete Fourier transform
DNN	Deep neural network
FE	Feature enhancement
GMM	Gaussian mixture model
HMM	Hidden Markov model
LVCSR	Large vocabulary continuous speech recognition
MDT	Missing data techniques
MFCC	Mel-frequency cepstral coefficient
NMD	Non-negative matrix deconvolution
NMF	Non-negative matrix factorisation
NSC, NNSC	Non-negative sparse classification/coding
PLP	Perceptual linear prediction/predictive
SC	Sparse classification/coding
SDR	Signal to distortion ratio
SE	Signal enhancement
SNR	Signal to noise ratio
STFT	Short-time Fourier transform
VAD	Voice activity detection

List of Included Publications

This thesis comprises the following publications, supplemented by introduction to the main topics and a summary of conducted research. Original publications are reprinted by permission from their respective copyright holders. The included publications are referred to in the text as [P1] . . . [P8].

[P1] A. Hurmalainen, J. F. Gemmeke, and T. Virtanen. "Non-negative Matrix Deconvolution in Noise Robust Speech Recognition". In: *Proceedings of the 36th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Prague, Czech Republic, 2011, pp. 4588–4591.

[P2] A. Hurmalainen, K. Mahkonen, J. F. Gemmeke, and T. Virtanen. "Exemplar-based Recognition of Speech in Highly Variable Noise". In: *Proceedings of the 1st International Workshop on Machine Listening in Multisource Environments (CHiME)*. Florence, Italy, 2011, pp. 1–5.

[P3] A. Hurmalainen and T. Virtanen. "Modelling Spectro-Temporal Dynamics in Factorisation-Based Noise-Robust Automatic Speech Recognition". In: *Proceedings of the 37th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Kyoto, Japan, 2012, pp. 4113–4116.

[P4] A. Hurmalainen, J. F. Gemmeke, and T. Virtanen. "Detection, Separation and Recognition of Speech From Continuous Signals Using Spectral Factorisation". In: *Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*. Bucharest, Romania, 2012, pp. 2649–2653.

[P5] A. Hurmalainen, R. Saeidi, and T. Virtanen. "Group Sparsity for Speaker Identity Discrimination in Factorisation-based Speech Recognition". In: *Proceedings of the 13th Annual Conference of International Speech Communication Association (INTERSPEECH)*. Portland, OR, USA, 2012, pp. 2138–2141.

[P6] A. Hurmalainen, J. F. Gemmeke, and T. Virtanen. "Modelling non-stationary noise with spectral factorisation in automatic speech recognition". In: *Computer Speech & Language* 27.3 (2013), pp. 763–779.

[P7] A. Hurmalainen and T. Virtanen. “Acquiring Variable Length Speech Bases for Factorisation-Based Noise Robust Speech Recognition”. In: *Proceedings of the 21st European Signal Processing Conference (EUSIPCO)*. Marrakech, Morocco, 2013, pp. 1495–1499.

[P8] A. Hurmalainen, J. F. Gemmeke, and T. Virtanen. “Compact Long Context Spectral Factorisation Models for Noise Robust Recognition of Medium Vocabulary Speech”. In: *Proceedings of the 2nd International Workshop on Machine Listening in Multisource Environments (CHiME)*. Vancouver, BC, Canada, 2013, pp. 13–18.

Author's Contributions to the Publications

The author of the thesis has acted as the primary contributor of each included publication, formulating the novel scientific content, producing the text and figures, and conducting the experiments. Exceptions consists of the following contributions by other authors:

- Tuomas Virtanen suggested applying the convolutive model of [P1] to speech recognition, studying the dynamic features of [P3], and using the group sparsity constraint of [P5]. Furthermore, he has supervised all research covered by the thesis and provided comments for the final appearance of each publication.
- Jort F. Gemmeke is the main author of the earlier AURORA-2 -based sparse classification framework, which was modified for experiments by the author in [P1], and formed the conceptual basis for later versions rewritten by the author. He has repeatedly assisted on conducting the experiments, contributed to textual content where listed as an author, and provided reference system results for [P1] and [P8].
- Katariina Mahkonen provided the atom-state mapping algorithms used repeatedly in sparse classification from [P2] onward.
- Rahim Saeidi conducted the speaker recognition experiments and provided the results in [P5].

Chapter 1

Introduction

1.1 Of Speech and Recognition

For humans, communication by speech is natural and so trivial that even a child can do it. For machines, it is neither natural nor trivial. Nevertheless, in a world where communication, and more generally all processing of information, is turning digital, there is high request for automated systems capable of handling speech. Applications of such systems include hands-free operation, speech-controlled services, automatic transcription, and audio data mining.

Automatic speech recognition (ASR) stands for conversion of captured speech waveforms into their linguistic content. In other words, an ASR system receives an analogue or digital representation of sound waves as its input, and attempts to interpret it as a sequence of lexical units conveyed by the speaker [133]. Typical ASR problems can be characterised by recognising a sequence of *words* belonging to a natural language. A word sequence of convenient length for processing and interpretation is called an *utterance*. Multiple consecutive utterances without artificial division between them form *continuous speech*. Each of these context lengths appears frequently in speech applications and research.

1.1.1 Attributes of Speech

In addition to its lexical content, the actual waveform produced by a human speaker incorporates further information and features, including a personal voice profile, dialect, stress, emotion, pace and clarity, among others. These factors become instantly apparent by attempting automatic inversion with *speech synthesis*. If no side information is provided, the synthetic output can be expected to be unnaturally mechanical, or simply put, inhuman.

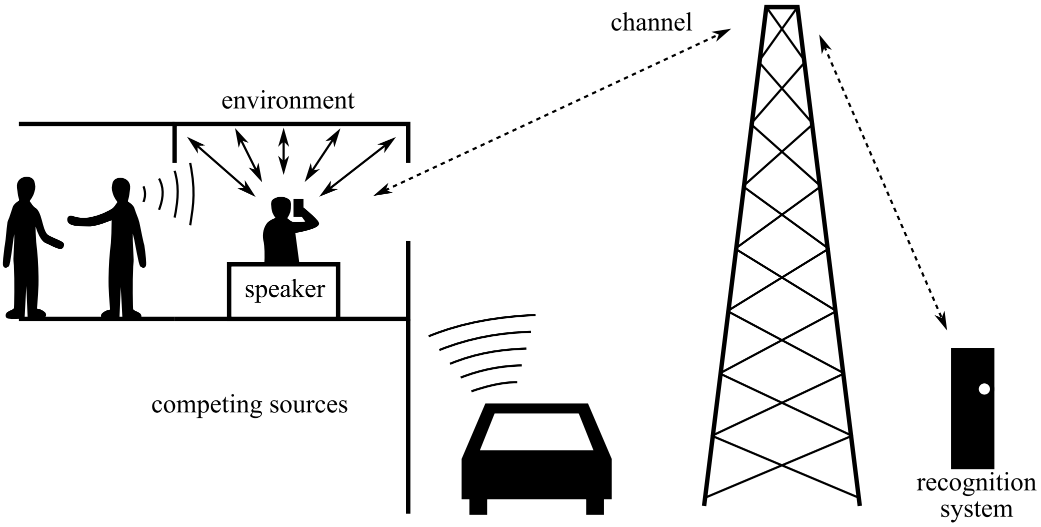
Figuring out the identity of a speaker forms a branch of its own, known as *speaker recognition* or *identification*. Other speaker traits and non-lexical features of speech are often bundled under umbrella terms *prosody* and *paralinguistics*. Stripping such parameters and only observing the lexical content leaves us with the problem of pure ASR. While two people reading the same utterance aloud may produce greatly differing waveforms, ideally the ASR system should still be able to translate both back into the same sequence of words. Therefore the defining function of ASR is to discover the fundamental content of speech from real-world utterances, which vary between speakers and unique instances. Although the other characteristics of speech may not be directly relevant for this goal, the system should be able either to take them into account or to mitigate them.

1.1.2 External Factors

Apart from variance introduced by the speaker, the observed waveform is further altered by factors, which we can roughly split into three categories: competing sources, the environment, and the transmission channel. The first comprises natural and artificial sound sources, ranging from wind to traffic, machinery, and competing speakers, whose voice should be suppressed instead of being recognised. While some of these sources could be considered a part of a certain *environment*, here we define it as a physical location, which introduces its characteristic acoustic phenomena like echo, reverberation and attenuation. Again, there is potential overlap with the *channel*. An acoustic path is always present until sound reaches a recording device, but in this work the channel is defined to cover electromechanical properties such as the microphone's response, signal bandwidth, distortion, compression, quantisation, and numerous types of transmission errors. The final waveform received by a recognition system will be affected by all of these factors to a lesser or greater extent.

Figure 1.1 shows an example of a scenario, where a remote ASR system is used over a phone link in a noisy office environment. When the speech signal finally reaches the recognition system, it has already been corrupted by room echo, noise sources, and a compressed, limited-bandwidth telephony channel. An ASR model trained from clean, close-talking speech is likely to encounter major problems, resulting in sub-par recognition rates. Another point of view to the same issues is given in Figure 1.2. A spectrogram of a short utterance is first shown as a clean recording, then under room reverberation, next in stationary white noise, and finally in non-stationary household noise containing a slamming sound, footsteps, and another human voice. Discovering the target speech features becomes increasingly difficult along the amount and complexity of interferences.

Figure 1.1: External factors contributing to real-world speech recognition. The signal received by a remote ASR system is corrupted by competing sound sources, room echo, and channel degradation.

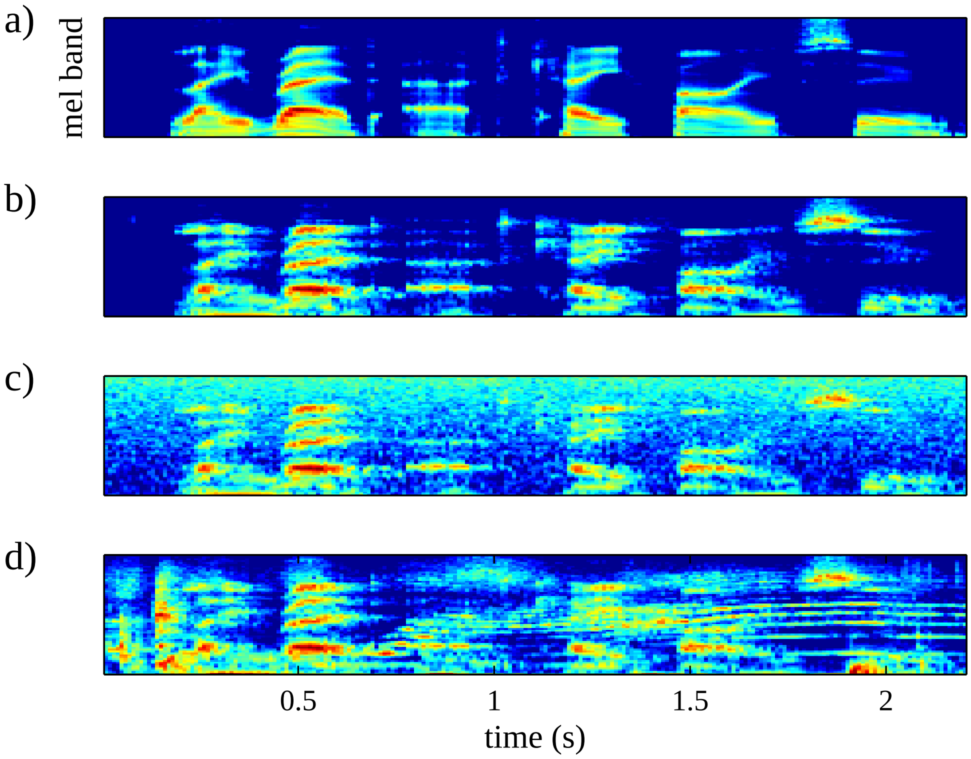


1.1.3 Current State of Machine Listening

Already from this brief description of speech production, sound transmission, and machine listening it becomes apparent that the signal received by an ASR system has no trivial one-to-one correspondence to its originating lexical content. What may appear a simple problem on a high level is in reality extremely complex — sometimes even impossible if any of the parameters in the overall scenario fall too far from their reasonably expectable range. Tackling the vast range of deviations and interferences is a major challenge, which in spite of decades of intensive research still tends to produce underwhelming results.

Although fluent machine listeners have been regularly envisioned in science fiction and future predictions, and actual systems have been developed since the 50s, only recently they have reached somewhat plausible reliability in a limited range of applications. Examples of ASR systems deployed for a wide audience can nowadays be found in voice-controlled personal assistants of mobile devices, automated captions of streamed video, in-car interfaces, instant translation, services over telephony, and games. While such systems can already prove helpful in many scenarios, an end user will regularly encounter recognition errors especially for casual speech and corrupted signals. 10–30% word error rates are still observed e.g. in voice search queries, while for the very diverse source material of online videos the failure rate may be around 50% [67].

Figure 1.2: Corruption of a speech spectrogram by external factors. Mel-spectral features of a short utterance are shown a) from the original close-talking recording, b) in room reverberation, c) in 0 dB SNR white noise, and d) in 0 dB SNR non-stationary real-world noise.



In practice, human hearing still constitutes state of the art for general listening tasks. This is not a complete surprise, considering how human speech production, natural languages and hearing have evolved in conjugation for millennia to facilitate efficient communication in a large variety of real-world environments. The processing power and adaptivity of brain for solving recognition problems should not be underestimated either. Human listeners also have broad knowledge on numerous real-world topics, the actual meanings of words, and non-verbal cues about the context, which together allow predicting the most likely message even from partially obscured or casually spoken utterances. Generic ASR systems lacking the same information start from deeply disadvantaged a position. For an uninformed ASR system, the often quoted semi-heterographs "recognise speech" and "wreck a nice beach" may appear deceivingly close, whereas for a human listener the latter is almost certainly nonsensical and instantly rejected. Especially in noisy environments the performance gap between human and machine listeners increases rapidly, and there is major interest in bridging this gap, not only in performance but also concerning the concepts involved in processing [150, 173].

1.1.4 Robust Processing

Current ASR systems reach acceptable accuracy for clearly articulated speech in good conditions. However, for widespread adoption of ASR systems for daily use they should be able to cope with significantly larger a variety of scenarios. On one hand, this requires better linguistic models for decoding casual utterances where dialects, slang, informal structures and mispronunciations are commonplace. On the other hand, higher robustness against external factors, especially competing sources, is needed. A multitude of robust speech processing methods have been proposed and discussed in literature. A thorough review of such methods is not viable for inclusion in this thesis, but there are books [85, 103, 107, 193, 207] and review articles [3, 50, 99] providing a broad overview to the field of separation, enhancement, compensation and robust recognition techniques.

An integral problem in robust ASR is the mismatch between models and observed speech. Its reduction is a recurring theme in all research directions. In the 1995 study by Gong [50], three main categories of methods were defined:

1. finding noise resistant features,
2. enhancement of noisy speech, and
3. compensation of speech models to match the noisy input.

All these directions are still prominently present in ASR research. More recently, Li et al. presented a more detailed taxonomy of robust methods using five different aspects for their categorisation [99].

To give a few examples of methods in each of Gong's categories, proposed robust features include decorrelated critical band energies such as mel-frequency cepstral coefficients (MFCCs), perceptual linear predictive (PLP) coefficients, relative spectral (RASTA) processing, Gabor filters, cepstral mean and variance normalisation (CMN, CVN), and strongly auditory-inspired features like power-normalised cepstral coefficients (PNCCs) and spectro-temporal receptive fields (STRFs) [162]. Alternatively, artificial neural networks (ANN) and especially their deep (DNN) versions can be used for acoustic modelling [67]. Enhancement methods include spectral subtraction (SS), noise masking, template processing, and explicit source separation with e.g. factorisation or spatial methods. For model compensation, parallel model combination (PMC), stereo-based piecewise linear compensation for environments (SPLICE), vector taylor series (VTS), maximum likelihood linear regression (MLLR), and uncertainty processing appear in literature [50, 99].

Regardless of other algorithm choices, the model mismatch can also be reduced simply by supplying matching training data, either real or artificially contaminated. However, at high noise levels this will result in models too wide for reliable recognition, thus it becomes necessary to apply an enhancing front-end to all data. Enhancement itself is also widely used in transmission, storage and further processing of speech, even if ASR is not involved.

The main focus of this thesis is on separation and recognition of speech from difficult mixtures, where several competing sound sources overlap the target speech. Unlike in many other robust schemes, few assumptions like stationarity are made on the noise. Key concepts for achieving this goal include *non-negative matrix factorisation* [12], *compositional models* [194], and *exemplar-based* methods [146], which have gained interest and produced strong results in robust speech processing within the last decade [87, 183]. Beside overall improvement in separation and recognition accuracy, the work aims at higher adaptivity to variable conditions and more efficient models for practically viable implementation. Published results and independent evaluations suggest that significant progress has been made in these areas in both standalone and combined systems.

1.1.5 Summary

To summarise this short introduction, the overall problem of real world speech recognition with its numerous issues and contributing factors is already on a theoretical level more difficult than it might first appear. The vast amounts of money and time spent on the problem with only partially acceptable results have proven it difficult in practice too. To the fundamental question of whether we are “there yet”, for certain tasks and conditions the answer could be a hesitant “maybe”, while for many other scenarios it would be “no”. Nevertheless, there is a reason to believe that algorithmic solutions can eventually match and even surpass the performance of human ear and brain. This work definitely makes no claims of such performance nor a complete solution to ASR, but it provides selected insights to novel methods, which demonstrably improve the robustness and adaptivity of ASR in conditions reflecting everyday situations.

1.2 Scope of the Thesis

The defining topic of this work is applying a family of *spectrogram factorisation* algorithms to speech data. The common goal is finding a composite model, which represents observed mixture spectrograms as an additive combination of more primitive speech and noise components. Because input signals are modelled explicitly with spectro-temporal events from multiple sources, speech content and other information can be extracted from complex mixtures of sound, which would be beyond the capabilities of models assuming a simpler interference profile. The work was motivated and directly follows early experiments on factorisation-based

ASR, which appeared to produce rapid increments in recognition rates especially for very noisy inputs [44, 190]. There is also considerable novelty in the approach as it blurs the line between separation and recognition that have typically been performed as separate steps in robust ASR systems [3]. More generally, it provides one potential answer to the request for new paradigms in ASR research, where more conventional routes have seen gradual improvements but no major breakthroughs, which would be welcome for expanding the scope of machine listening.

The speech processing systems of this thesis cover three broad stages:

1. selecting the features and models for sources,
2. factorisation of observed spectral data, typically noisy speech, and
3. using the factorisation output for separation and recognition.

Novel and refined methods are presented for each of these stages. Special attention is also given to model adaptivity and computational complexity, which in earlier work have turned out as challenges for practical applicability of the approach [44]. Experiments are conducted using a factorisation and recognition framework developed at Tampere University of Technology in close collaboration with other institutions. Results are evaluated using publicly available speech databases with objective metrics including speech and speaker recognition accuracy, and separation quality.

1.3 Scientific Contributions

Main scientific contributions of the thesis comprise:

- Refinement and extension of mel-spectral feature spaces for higher separation and classification capability [P3, P6].
- Advancing from randomly sampled, exemplar-based speech models to semi-random exemplars [P2], compact template models [P6], and clustered variable-length modelling for small and medium vocabulary [P7, P8].
- Acquiring noise models from the nearby context of target utterances [P2], directly from mixtures [P6, P8], adaptively from continuous inputs [P4], and by learning characteristic patterns from separate training data [P8].
- Using multiple speaker-dependent speech bases for speaker recognition [P5], closed-set identity estimation [P4], and approximation of new speakers [P8].
- New factorisation algorithm variants in the context of speech recognition, including convolutive modelling [P1], continuous input processing [P4], semi-supervised factorisation [P6], group sparsity for multiple speaker models [P5], and variable-length dictionaries [P7, P8].

1.4 Recap of Included Publications

Eight publications, listed in the front matter, are included in the thesis and summarised in the following list. Chronological order of publishing is mostly used with the exception of [P7] and [P8], which are switched for better reflecting the overall system development.

[P1] Non-negative Matrix Deconvolution in Noise Robust Speech Recognition

Non-negative matrix deconvolution (NMD) — also known as convolutive non-negative matrix factorisation (CNMF) — is proposed for modelling continuity in utterances instead of previously used factorisation of independent, overlapping windows. Although convolutive NMF had been previously demonstrated for tasks such as two-speaker separation [158], this is the first known system performing speech recognition via sparse classification (SC) convolutively.

[P2] Exemplar-based Recognition of Speech in Highly Variable Noise

The paper describes a factorisation and recognition framework developed especially for 1st CHiME Challenge data [3]. Noisy speech recognition performance is evaluated using multiple algorithm variants with the overall system submitted as a challenge entry in a dedicated workshop. The proposed system introduces deriving noise models solely from the neighbouring context of utterances, and modelling speech with partially informed exemplar selection. Previously published concepts of convolutive modelling [P1] and learnt activation-state mapping [106] are employed in the challenge system.

[P3] Modelling Spectro-Temporal Dynamics in Factorisation-Based Noise-Robust Automatic Speech Recognition

In this paper, mel-spectral feature spaces are studied further within the factorisation framework. First, the number of spectral bands and different band weighting methods are compared regarding ASR quality using static spectral features. Second, spectro-temporal dynamics are added to feature vectors by using Gabor and delta filters. Experiments show that separation and recognition quality can be improved by modelling temporal dynamics in NMF.

[P4] Detection, Separation and Recognition of Speech From Continuous Signals Using Spectral Factorisation

Whereas earlier experiments on the main framework assumed perfect knowledge of active speakers and temporal locations of speech and noise segments, the presented system makes considerably fewer assumptions. Long inputs are processed by estimating speech presence with NMF algorithms, and noise models are updated continuously from segments appearing noise-only, that is, devoid of target speech. Furthermore, it is compared how accurately the identity of an unknown speaker can be estimated within a closed set, and how much accuracy is lost by using an estimated speaker model instead of known identity. Overall, the work aims at more realistic processing of unannotated real-world inputs, where utterances and speakers may appear unpredictably, and noise models must be adapted on the fly to match the observed noise types.

[P5] Group Sparsity for Speaker Identity Discrimination in Factorisation-based Speech Recognition

The work extends previously published factorisation-based speech and speaker recognition methods by introducing a group sparsity criterion for multiple speaker-dependent models, which are used to estimate the best matching model for unknown speakers. Even though already the earlier system managed to produce reasonably accurate identity estimates and speech model approximations, group sparsity is found to sharpen the distribution of speaker candidates, thus making the final model more accurate for both speech and speaker recognition in noisy conditions.

[P6] Modelling non-stationary noise with spectral factorisation in automatic speech recognition

The article in a “special issue on speech separation and recognition in multisource environments” wraps up the authors’ best-performing separation and sparse classification systems on 1st CHiME Challenge data. In addition, new algorithms are proposed for speech and noise modelling, including informed acquisition of compact speech template models, and semi-supervised learning of noise models directly from observed mixtures. The new variants reduce model sizes and computational complexity significantly, thus making the approach more viable for practical implementations with limited memory and computational power.

[P7] Acquiring Variable Length Speech Bases for Factorisation-Based Noise Robust Speech Recognition

The paper proposes modelling speech with variable-length atoms, capable of addressing the large variance of phonetic units that appear in real-world speech. Beside describing the variable-length factorisation model, algorithms are given for acquiring speech bases by observing recurring units in the spectral domain, in state transcriptions, and combining both. Speech model sizes are reduced in comparison to the template bases introduced in [P6], while separation and classification quality is either retained or improved.

[P8] Compact Long Context Spectral Factorisation Models for Noise Robust Recognition of Medium Vocabulary Speech

The contribution to the 2$^{\text{nd}}$ CHiME Challenge and workshop combines several new methods, such as multiple speaker models, group sparsity, variable-length modelling, and semi-supervised noise modelling. Methods are applied to the noisy medium-vocabulary Track 2 corpus of the 2013 CHiME challenge. Compared to AURORA-2 and 1$^{\text{st}}$ CHiME corpora, this task reflects more closely real-world ASR problems. Proposed methods have been selected accordingly with the goal of providing practically applicable tools for tackling such problems, including model acquisition and computational complexity.

1.5 Organisation of the Thesis

The rest of the thesis is organised followingly. Chapter 2 introduces the fundamental concept of spectrogram factorisation, its motivation, and applications. An overview is given to the model and prior work on it, especially in the context of speech processing. Chapter 3 presents deriving feature spaces and source models, which are employed in factorisation. Actual factorisation algorithms and their variants are discussed in Chapter 4. In Chapter 5 it is shown how the output is used for speech recognition and related tasks via enhancement, classification, and joint techniques. Chapter 6 summarises results from evaluations and discusses practical issues of factorisation-based recognition. Finally, conclusions and discussion on further research topics are given in Chapter 7.

Chapter 2

Spectrogram Factorisation

2.1 Concept and Motivation

The spectrogram factorisation algorithms covered by this thesis are based on *non-negative matrix factorisation* (NMF), which is essentially an algorithm for finding a *composite model* for an observed mixture using a combination of lower-level components [194]. The main assumption is that individual sound sources have their characteristic spectro-temporal patterns, which can be modelled with a *basis* or a *dictionary* of features. Due to recurrence of patterns, a source can be modelled with less overall data than e.g. the whole set of its training spectrograms. Furthermore, assuming and enforcing such structure on source models makes it viable to separate multiple sources from a mixture with fewer channels than sources, which would otherwise be an ill-posed problem.

A central concept in the approach is strict *additivity*, that is, approximating the observation with a combination of basis components with only *non-negative weights*. The motivation is that despite occasional cancellation in the time domain, concurrent signals are additive in the spectral domain in the expectation sense. Especially for spectro-temporally sparse signals, where the energy of each source is concentrated on a limited amount of spectrogram bins, the assumption of additivity has been found to hold sufficiently well in practice too [189]. Therefore a non-negative model yields better estimates of contained sources than allowing negative weights so that source models could get subtracted from each other in the spectral magnitude domain.

Another concept frequently appearing in literature of the field and in this thesis is *sparsity*. Like already postulated by William of Ockham, “plurality is not to be posited without necessity”. The principle, better known as Occam’s razor, suggests that a simple model should be preferred over more complex alternatives.

Typical examples of the principle in the behaviour of real-world audio signals include that

- sound events are localised in the spectro-temporal domain,
- characteristic patterns of a single source belong to a limited set of all possible spectrograms, and
- the number of concurrent sources in a mixture is limited.

Therefore sparsity constraints are set on the model to favour solutions, where some or all of these assumptions are met. It will be later seen how sparse models lead to improvements in separation and classification.

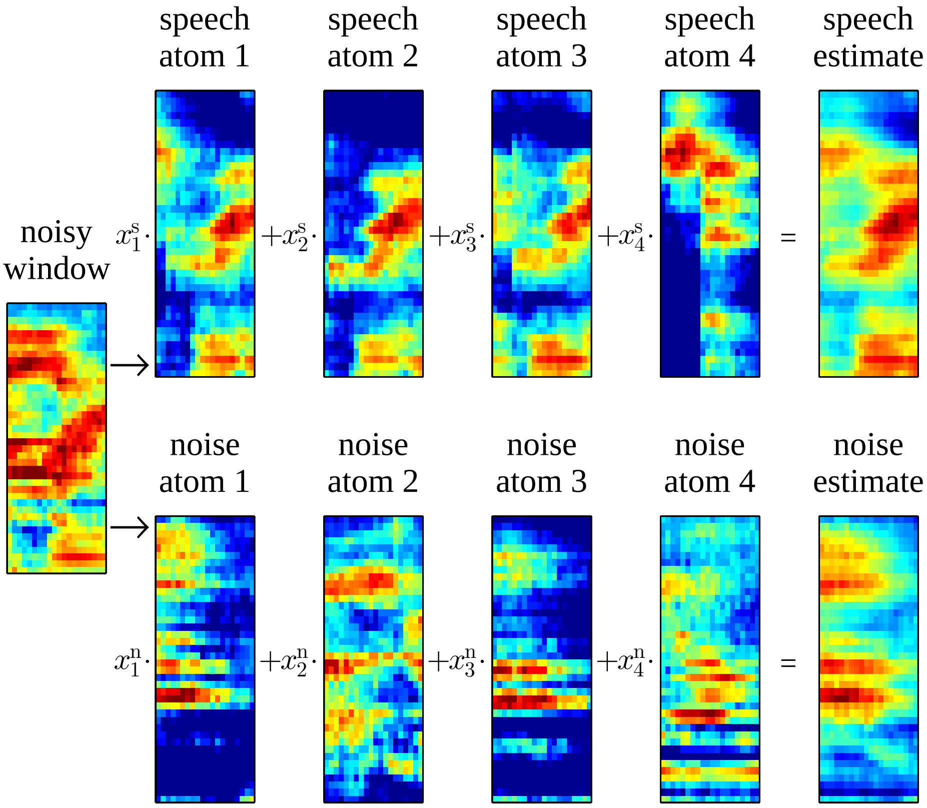
2.2 Applications

Finding a plausible estimate of the presence of underlying sources in the spectro-temporal domain instantly facilitates constructing single-source spectrogram estimates or a time-varying filter picking desired sources from the mixture. An example is shown in Figure 2.1, where a noisy spectrogram window is modelled as a weighted sum of four speech and noise atoms, which together produce speech and noise spectrogram estimates. The process is described in more detail in Chapter 5. By reconstructing phase information not present in the spectral magnitude domain, source estimates can be brought back into the time domain. These techniques are referred to as *source separation*, *feature enhancement* (FE), and *signal enhancement* (SE). In music applications, the techniques are used for selecting or suppressing specific instruments from mixed tracks [62, 81]. In computational audio scene analysis (CASA), sound events of interest can similarly be separated from multi-source scenes.

In speech processing, *separation* stands for e.g. splitting two speakers' voices apart from a mixture, whereas *enhancement* means improving the objective quality or intelligibility of a selected voice among other sources, which together are called *noise*. In practice, however, the problems are tightly related and often solved with similar algorithms.

Although feature or signal level separation is arguably the most commonly employed application for NMF algorithms in the context of audio signals, another notable branch is using sparse decompositions for *classification*. In a sense, already the previously described separation methods function as classifiers, because they select components belonging to single-source sets, which are then separated. However, further information can be extracted by observing *which* characteristic patterns are detected in the input and *when*. In the example seen in Figure 2.1, the information for classification is derived from the weight coefficients x by exploiting knowledge of audio events and phonetic patterns present in each atom.

Figure 2.1: Approximating a noisy spectrogram window as a weighted sum of speech and noise atoms. Finding the optimal weights x for each basis atom produces speech and noise spectrogram estimates, while also revealing more primitive audio components most likely present in the observed mixture.



In music applications, sparse and non-negative algorithms known as *sparse classification* or *sparse coding* (SC) have been used for instrument recognition [142], genre classification [126], and automatic transcription [6, 90, 159]. Similarly in audio scene analysis sound sources can be identified by SC [15, 49, 61]. In speech processing, SC has been used for ASR [37, 44] and speaker recognition [74, 143, 177]. These speech applications form a major topic in this thesis.

Finally, we can define a third important application for NMF algorithms, namely *model learning*. In many variants of NMF, the system not only classifies signal components, but also learns some or all of its models with a factorisation algorithm either beforehand or online. Especially sparsity and assumption of redundancy play a major role in finding compact models for potentially large source data sets. The topic of NMF-based model learning will come up repeatedly in this work.

2.3 Fundamental NMF Model

The core NMF model can be summarised by equation

$$\mathbf{y} \approx \sum_{l=1}^{L} \mathbf{a}_l x_l \tag{2.1}$$

or equivalently as a matrix-vector multiplication

$$\mathbf{y} \approx \mathbf{A}\mathbf{x}. \tag{2.2}$$

Here $\mathbf{y}$ is the *observation vector* being modelled. It is a length B column vector, where B is the dimensionality of our feature space. For spectrogram processing, B is the number of *spectral bands*, hence the symbol choice. The observation is approximated by a weighted sum of *atom vectors* $\mathbf{a}_l$, also belonging to $\Re_{\geq 0}^{B\times 1}$. Atoms are indexed by $l \in [1, L]$, where L is the total number of atoms in the *basis* or *dictionary* $\mathbf{A}$ $(B \times L)$. Vector $\mathbf{x} \in \Re_{\geq 0}^{L\times 1}$ contains the *activation weights* of each atom. Importantly, all data vectors and weight coefficients are assumed strictly *non-negative*. The estimate of observation $\mathbf{y}$ is denoted by $\boldsymbol{\psi}$ (or occasionally $\tilde{\mathbf{y}}$).

Only the observation vector $\mathbf{y}$ is always assumed externally given and thus a constant. With no other constraints, equation (2.2) is obviously underdetermined, because a single basis vector $\mathbf{a} = \mathbf{y}$ with a trivial weight 1 will give a perfect spectral estimate, and any further basis vectors would be redundant. However, under additional constraints or in other formulations, the triviality no longer applies.

2.3.1 Matrix Form

First, the same basis is commonly used to model multiple observations. For now, let us denote the number of observation vectors by I. The vectors are conventionally gathered to an *observation matrix* $\mathbf{Y} \in \Re_{\geq 0}^{B\times I}$, and estimated by matrix $\boldsymbol{\Psi}$ (or $\tilde{\mathbf{Y}}$) of the same size. Similarly, each observation vector will have its own activation vector in an *activation matrix* $\mathbf{X} \in \Re_{\geq 0}^{L\times I}$ so that

$$\mathbf{Y} \approx \mathbf{A}\mathbf{X}. \tag{2.3}$$

If $L < I$, in a general case the problem becomes overdetermined, and solving it will actually find a compressed approximation of $\mathbf{Y}$ with a reduced number of components, potentially revealing something about its underlying structure. This formulation is arguably the most common version of NMF and the core of the seminal work by Lee and Seung [91, 92].

2.3.2 Levels of Supervision

As a second constraint on the problem, parts of the basis $\mathbf{A}$ may be fixed. If the content of $\mathbf{A}$ is wholly adapted during factorisation, the problem is called *unsupervised*. If some of $\mathbf{A}$'s vectors are fixed, we call the problem *semi-supervised*. If $\mathbf{A}$ is completely fixed beforehand, the factorisation is called *supervised*. In the latter case, we try to find the best non-negative combination of given basis vectors to approximate the observation. An exact solution exists if and only if $\mathbf{y}$ lies within the cone spanned by the basis vectors with non-negative weights. Furthermore, it can be shown that up to B atoms suffice for finding the best solution regarding the residual [191]. Potentially, albeit rarely, (semi-)supervised variants are formulated by fixing $\mathbf{X}$ and learning $\mathbf{A}$. One such example can be found in [70].

2.3.3 Spectral Distance Measures

The primary target in modelling is to minimise a distance function between the observation vector $\mathbf{y}$ and its estimate $\boldsymbol{\psi}$ or their matrix counterparts. One common choice is (squared) Euclidean distance,

$$d_{\mathrm{Euc}}(\mathbf{y}, \boldsymbol{\psi}) = \sum_{b=1}^{B} (y_b - \psi_b)^2. \tag{2.4}$$

However, as this measure is often dominated by the largest elements in the vectors, it has been found more appropriate to use another measure, which emphasises differences in small-magnitude bins. In audio spectrogram applications and also this thesis, a common alternative is *generalised Kullback-Leibler* (KL) *divergence*,

$$d_{\mathrm{KL}}(\mathbf{y}, \boldsymbol{\psi}) = \sum_{b=1}^{B} y_b \log \frac{y_b}{\psi_b} - y_b + \psi_b. \tag{2.5}$$

While not a distance function in a strict sense due to its lack of symmetry, the measure is applicable and commonly employed in audio literature [189]. Other alternatives have been proposed and compared e.g. in [10, 11, 30, 31, 94, 161, 211], and a comprehensive list of distance function can be found in [12].

2.3.4 Sparsity

In classification and separation applications, it is commonplace to employ *sparsity cost functions* to control the structure of $\mathbf{x}$ or $\mathbf{X}$. In sparse solutions, a majority of total activation weight is condensed on relatively small a number of atom indices with the rest being zeros or negligibly small. Sparsity is especially integral for supervised factorisation tasks employing overcomplete bases. A distance cost function like (2.4) or (2.5) often has infinitely many equivalent solutions, because any linearly dependent vectors in them are equally interchangeable to another set of vectors, and nothing in the distance measure is limiting the complexity of the activation pattern. Conversely, enforcing sparsity will find a small number of best matching basis vectors to explain the observation, producing more plausible results in many separation and classification tasks according to Occam's principle. A small mismatch in the spectral estimate is acceptable if the underlying composite model can be simplified to better reflect the behaviour of real-world sources.

The most straightforward definition of sparsity known as the L_0 norm — the number of nonzero elements in $\mathbf{x}$ — could be viable, but it often leads to NP-hard optimisation problems. Also, for our purposes the difference between strict zero and near-zero weights is of no particular importance. More commonly the measure is replaced with L_1 norm, that is, the sum of all activation coefficients in $\mathbf{x}$, which makes the optimisation problem convex again. However, other cost functions have been proposed for general sparsity [76] or for special purposes like *group sparsity*, discussed in Section 4.3 [P5, 165].

2.3.5 Solving NMF

No closed form solution is known for finding the optimum of common NMF problems, not even for the convex supervised case. Therefore minimisation is performed with iterative descent algorithms, which can be found in literature for common cost functions [12, 91, 92]. While implementations of these algorithms do not belong to the main scope of this thesis, their behaviour must be taken into account occasionally. From a practical point of view, a persistent issue is computational complexity, which can grow high for large data sets and iteration counts. The problem of complexity has been tackled on one hand by developing faster descent algorithms, and on the other by finding more compact models for the data. A few lines of ongoing research on these topics are given in Chapter 6.

2.3.6 Extensions

The basic NMF model only factors single observation vectors or groups of vectors, treating them as disjoint units. In audio applications, however, a lot of information lies in temporal continuity of signals over consecutive spectrogram frames. Thereby it is highly beneficial to extend the model to support multi-frame *spectro-temporal patterns* and their correlations, e.g. by using continuity constraints [189], a larger window size [47, 157, 158], or priors for transitions between consecutive frames [53, 119]. These extensions are discussed in more detail in Section 4.1. Gradually, several other priors [51] and structures [12] have been proposed for NMF, including tensor factorisation [33], source-filter models [192], orthogonal NMF [23], semi-negative factorisation [22], convex hull NMF

[174], GMM priors [55] and so on. However, these and other advanced variants are mostly beyond the scope of this brief introduction and the work covered by the thesis. A few of these extensions are discussed later as potential future directions.

2.4 Prior Work in Speech Processing

2.4.1 Early NMF: From Music to Speech

Since its introduction, NMF has found its way into signal processing [69, 102], then more specifically to audio via separation of music [4, 62, 152, 159, 187, 188, 195], and finally to speech applications [153]. Unlike typical music and audio scenes, which are inherently composite mixtures, speech is a single-source process and employing a factorisation algorithm might first appear unnecessary. However, when we consider real-world applications, where overlapping speakers and noise sources are the norm, it becomes highly relevant again to construct single-source models, which are then employed in NMF to separate desired voices from mixtures. In addition, NMF has the capability to perform *sparse approximation*, where a slightly differing instance of a sound is represented as a sparse combination of nearest training instances. The large variation in casual pronunciation makes it unlikely to find a perfect match to observed speech, thus joint approximation with multiple candidates is well motivated in speech modelling.

2.4.2 Enhancement and Missing Data Masks

Early speech applications of NMF emerged in 2006, when Schmidt and Olsson proposed applying basic NMF techniques to two-speaker separation [153]. In their work, speaker-dependent single-frame models were learnt by factorisation of either a set of utterances as a whole, or by segmenting the training corpus into single phonemes. Then conventional supervised NMF with two speakers' bases was applied to a mixture. Already here sparsity was found significant for separation performance. The model was soon extended with convolutive multi-frame modelling by Smaragdis [158]. A comprehensive study of basis parameters and separation results was provided for two-speaker and speech-noise mixtures. Thereafter NMF enhancement has become an established method in robust ASR [73, 118, 135, 161, 185, 203, 206, 208].

Apart from separation and enhancement, spectrogram estimates and atom activations have been used for *uncertainty mask estimation* and *imputation* [38, 39, 44, 80] in recognition based on *missing data techniques* (MDT) [13, 85, 136]. In this approach, factorisation output is used to assign a certainty measure to spectrogram bins, depending on how reliably they are expected to represent actual speech.

Missing and unreliable features are either reconstructed or given reduced weight in decoding. Several methods for estimating certainty are proposed in literature [78, 79]. However, they are not employed in this thesis.

2.4.3 Sparse Classification

An alternative route dubbed *sparse classification* or *sparse coding* (SC) was proposed for speech recognition in 2008, first for simplified single-digit [37] and multi-digit recognition tasks [47]. In the first example, fixed-length multi-frame feature vectors were constructed for digits by interpolation. For the multi-digit case, a *sliding window* approach was introduced, modelling variable-length utterances as a set of overlapping multi-frame windows. In sparse classification, each atom is given identifiers or *labels* describing its content such as phones, states, or word membership, whereafter recognition is performed by observing the activation weights of labelled atoms. Later the main approach has been extended by frame-level labelling of atoms [190] and more advanced label learning algorithms [70, 106]. Other applications for SC in speech processing include speaker identification [P5, 86, 120, 143], age and gender estimation [2], speech overlap detection [34, 186], and detection of non-linguistic vocalisations [154].

2.4.4 Exemplar Models

A prominent concept employed in compositional modelling of real-world sounds is using *exemplars* as models for speech and noise [146]. Arguably the most commonly used approach for spectral model acquisition in NMF literature has been learning from a training corpus using single-frame or convolutive multi-frame models [153, 158]. These are expected to capture a compressed model of characteristic patterns of speech in relatively small a number of atoms. In exemplar modelling, atoms are sampled directly from spectrograms of a training corpus or the neighbouring context, forming an overcomplete basis of specific instances of acoustic patterns. The expectation is that an observed pattern can be approximated as a linear combination of exemplars representing similar events. Natural labelling follows from the fact that each speech atom is picked from a segment of speech, whose lexical and phonetic content is typically known.

Outside NMF, the use of exemplars or templates as classifiers goes a long way back to the early days of ASR [18, 66, 97, 134]. Refined variants have later been proposed by several groups [17, 21, 60, 145, 164, 166]. Compared to plain template matching, the novelty of exemplar-based NMF lies in its additive model, permitting joint approximation with multiple, overlapping models for one or more sources simultaneously. Exemplar models have later found their way into speaker recognition [143] and voice conversion [171, 210].

Chapter 3

Features and Source Models

A fundamental stage in any recognition task is defining the features and models, which facilitate later detection and classification. This also applies to source separation, which generally is an ill-posed problem when the number of sources exceeds the number of input channels. Therefore its success rate heavily depends on the *source models*, which form a constrained subset of all possible spectro-temporal patterns. A well-defined model is expected to match its corresponding source, rather than any other sounds present in the mixture. Models and observations are processed within a spectro-temporal *feature space*, which is defined for the task taking into account e.g. accuracy, robustness, model size, and computational costs. This chapter describes feature spaces and model acquisition methods, which form the basis for later factorisation steps.

3.1 Features Spaces

3.1.1 Spectrogram Representation

Digital audio signals are generally recorded as time-domain pulse-code modulation (PCM) waveforms. However, in analysis of natural sounds it is commonplace to transform the inputs into a spectral domain, where periodic components from resonating sources, such as the vocal tract, become concentrated into spectral bins corresponding to their frequency. A time-frequency representation, where energies of spectral bands can be observed over time is called a *spectrogram*. It is most commonly acquired by *Fourier analysis* of short-term *frames*.

Especially in single-channel applications it is common to discard phase information, which corresponds to the complex angle of spectral domain coefficients. Only signal energy is observed, which will be invariant to the exact phase alignment of the originating time-domain signal. For NMF algorithms, this has the immediate benefit of producing strictly non-negative real values for the spectral content of observed mixtures and basis components, which is required by standard NMF models. While most spectrogram processing like factorisation can be conducted using absolute values of spectral coefficients, reconstructing the phase information is still required for producing time-domain signals like enhanced speech. These methods are discussed later in Section 5.1.

Despite gradual introduction and adoption of multi-resolution transforms like the constant-Q transform or wavelets, a majority of spectral processing still takes place using spectrograms derived from fixed-length frames. The temporal resolution of the system can be characterised by two parameters, *frame length* and *frame shift*. The former defines the duration of each frame. It is set to a value where the input can be assumed mostly stationary, which for speech may stand for 20–64 milliseconds. Frame shift is the amount of input time advanced before extracting a new frame, and is usually set to 50% or less of frame length. The difference of these two values is *frame overlap*. In this work, frame length is uniformly 25 ms and the frame shift 10 ms, equivalently to common speech recognition back-ends and baseline models provided for evaluation tasks [3, 183, 213]. In ASR, the 100 Hz frame rate is regularly used for capturing rapid dynamics of speech, which helps in recognition. Conversely, longer frames are often used in signal enhancement, where slower transitions reduce audible artifacts from estimation errors. A few frame lengths seen in literature include 32 ms [197], 40 ms [135, 189] and 64 ms [137, 208] with values up to 256 ms compared in [158].

Time-domain frames are conventionally multiplied by a *window function* such as Hann or Hamming window, which will reduce edge artifacts caused by cyclic discontinuity in the following short-time Fourier transform (STFT) step. The STFT output, after discarding angle, is called the *magnitude spectrum* and their sequence over time the *magnitude spectrogram*. Bin-wise squaring of coefficients produces the *energy spectrum* and *spectrogram*, correspondingly.

3.1.2 Base Spectral Features

In NMF-based audio processing, the preceding steps are typically applied quite uniformly with differences arising mostly from input sample rate and frame length parameters. However, from this point on, a few conceptually different paths diverge. The foremost choices are

1. using spectral magnitudes or energies and
2. whether to apply a filter bank or not.

The magnitude versus energy choice, in conjugation with the NMF distance measure, affects how much large energy concentrations dominate the factorisation, and how well the assumption of approximate additivity will hold. In this thesis' publications, magnitude features are used according to earlier work, where spectral magnitudes have been found efficient for audio separation and enhancement tasks [44, 55, 189, 208].

More variation can be observed in filter bank use. In general purpose and music enhancement, the STFT magnitude or energy spectrum is often used by itself, providing the highest spectral resolution achievable for the chosen frame length. Conversely, in speech processing it is customary to employ an auditory-motivated filter bank such as mel, Bark or ERB compression. These nonlinear spectral band mappings derived from critical bands of human hearing aim at capturing the formants of voice at a sufficient resolution for classification, while removing exact pitch contours produced by the speaker's voice and prosody. In clean speech recognition, mel features, further compressed by mel-cepstral analysis, have been found viable for constructing speaker-independent systems robust against slight variations in the voice profile [133].

In NMF separation systems there are arguments both for and against critical band filter banks. Full spectral resolution will potentially yield maximal separation accuracy. However, it sets higher requirements on basis acquisition, because there will be potentially a large mismatch between e.g. instances of phones uttered at slightly different fundamental frequencies. Compressed filter bank features cannot separate events overlapping within the same band, but they are more robust against individual variation in articulation and external sound events. Major savings are also achieved in memory and computation costs, because the number of spectral bands to be processed is reduced from hundreds or even thousands to less than 50 in typical implementations. Assuming that a majority of spectral information crucial for speech processing can be compressed into a few bands, more resources can in turn be spent on other system parameters such as basis size.

Whereas conventional recognisers commonly employ cepstral transformation for deriving their features, in NMF systems it is avoided due to increased violation of additivity from logarithmic compression, and the introduction of negative coefficients from a cosine transform. Therefore the feature extraction process is stopped after deriving the magnitudes or energies in the spectral representation with optional reduction of spectral resolution by a filter bank.

Objective comparisons listed in Chapter 6 suggest that the proposed framework employing mel-filtered magnitude features is efficient for robust ASR, compared to systems with higher spectral resolution but otherwise smaller models. Nevertheless, the trade-off between spectral resolution, model acquisition methods, and output quality should be studied further. In [P3], different flavours of mel-spectral features were briefly investigated regarding the band count and their weighting schemes. Other work on optimisation of purely spectral features in the context of NMF is scarce, making it a potential topic for future research. Alternative filter banks such as gammatones should also be studied. Linear prediction based features like PLPs [63] often appearing in ASR systems cannot be ruled out either.

3.1.3 Multi-Frame Windows

As stated in the beginning of this chapter, the success rate of underdetermined separation depends on the system's ability to model underlying sources discriminately. Similarity of sources leads to uncertainty on the solution. The problem arises frequently in sound separation, because there may be significant overlap in the short-term spectral profiles of different sources.

A partial alleviation is achieved by explicitly modelling the spectro-temporal behaviour of sources in atoms and observation windows with extended context. While two sources may appear similar in a single pair of frames, observing their long-term behaviour is more likely to bring out the characteristic patterns, where the confusion no longer applies. The principle was already employed in early template matching algorithms, and reformulated as TRAPs features in 1998 [65] and their refined versions in 2003 [8]. Related extensions of the temporal context have been proposed via trajectory modelling [59, 214, 217] and neural networks [9, 67, 113]. Especially deep neural networks (DNNs) have rapidly gained popularity in alternative acoustic modelling. In a typical DNN system, the input block may comprise 7–37 frames from the base features such as mel-filtered or PLP representation, with or without decorrelating transformation [67]. Even longer windows have been used in neural network modelling, although it may be recommendable to reduce the dimensionality with principal component analysis (PCA) or linear discriminant analysis (LDA) to form the actual input layer [9].

In NMF separation, a larger window size has been proposed for both matrix deconvolution [157, 188] and sliding window processing [47, 52]. In either algorithm, a basis atom is expanded from a $B \times 1$ vector to a $B \times T$ matrix, which models the spectra of T consecutive frames. Note that in this work the term 'window' generally refers to a multi-frame spectrogram segment, even though in other audio signal processing it often overlaps with the concept of a 'frame', and in some cases such as 'window functions' this convention is still followed.

The amount of physical time being modelled depends on frame parameters and window size. In this thesis' publications, the window size ranges from 8 to 50 frames, standing for approximately 80–500 milliseconds of context. In earlier work, 5–30 frame windows were evaluated using multiple recognition methods [44]. Typically the best results were achieved with 200–300 ms windows, which already capture strongly discriminative spectro-temporal patterns [P1, P6, 44]. Other multi-frame window sizes found in literature include 70 ms [185], 80 ms [198], 176 ms [158], 224 ms [122] and 256 ms [203, 208] with results favouring the longer end. Increasing the size even further is obviously possible. However, it makes the atoms highly specialised, thus requiring a large basis in order to provide a good match to each observed instance of events. The issue naturally affects large vocabulary recognition and diverse noise events more than e.g. small vocabulary recognition tasks. Consequently long windows require plentiful training material, and they increase the computational costs of factorisation. Selection of frame and window length parameters is therefore task-dependent, and a compromise between quality, cost, and source data factors.

In later work, using multiple window lengths in parallel [212] and a mixed-length basis [P7] have been proposed for solving the trade-off. It is also possible to circumvent the rigidity of atoms by using histograms [178], interpolation [37], or HMMs [53, 115, 119]. Different temporal models are discussed concerning their basis acquisition in this chapter, and from a factorisation viewpoint in Section 4.1.

3.1.4 Dynamic Features

Physiological studies have shown that in addition to absolute energies in spectral bands, the human auditory system is specially sensitive to *spectro-temporal dynamics*, that is, transients of energies over time and frequency [130, 162]. There is an emphasised response in the firing rate of nerve cells to the on- and offsets of tones. This behaviour is regularly modelled by augmenting static features with time derivatives or *deltas*. Because improvements are commonly observed in recognition accuracy, delta features are implemented and available in common ASR software packages [213]. Beside first degree temporal deltas, second and even third derivatives are commonly used to capture temporal dynamics.

However, temporal dynamics only address transients in one spectrographic direction. Further studies on the auditory cortex have revealed specialised responses to complex spectro-temporal patterns and modulations [27, 110, 128]. Consequently there have been several attempts of finding more efficient pattern detectors using e.g. Gabor filter banks [83, 111, 112, 218], fuzzy logic units [84], cortical models [109], and learnt spectro-temporal filters [98]. Advanced spectro-temporal modelling can be expected to surpass the performance of more common static-only or delta-augmented features.

In the NMF approach, dynamic filter banks are less commonly used. One reason is that filtering tends to produce both positive and negative coefficients, which cannot be employed directly in a conventional non-negative framework. Another reason is that long multi-frame windows already capture spectro-temporal behaviour of events, which is the primary argument for augmenting single-frame features with temporal or two-dimensional filters in the first place. Moreover, large filter banks rapidly increase the feature vector length, commonly by an order or two of magnitude, making them impractical for large-scale NMF. Nevertheless, using simple delta filters will emphasise the on- and offsets also in NMF, thus incorporating some auditory-like processing in the system.

In [180], deltas were computed to the immediate neighbouring bins in spectral and temporal directions. Negative values were handled by splitting the filtered vectors into their positive and negative components. The same fundamental principle was used in [P3], where simple Gabor and delta filters were applied to basis and observation spectrograms. Probably due to the low resolution of mel spectra, no significant gains were observed from filters functioning in the spectral direction. However, modelling temporal dynamics was found beneficial. A temporal delta filter similar to common MFCC back-end processing [213] improved separation and recognition rates, hence it was also included in the best performing exemplar system of [P6].

3.1.5 Stereo and Multi-Channel Features

For a long time, ASR research concentrated primarily on optimising the clean speech recognition rate in monaural recordings such as telephone or single microphone applications. Conversely, in last decades there has been increasing interest toward multi-source scenarios with spatial dimensions. As multi-microphone devices and recordings are gradually becoming more widespread, spatial audio processing has become a major branch in separation and robust algorithm research.

Straightforward spatial features can be added to the basic NMF model by extracting features for each channel separately, and then concatenating the channel vectors. This was the approach used in [P2], motivated by the 1st CHiME Challenge corpus where binaural recordings are available. However, in [P3], the improvements in recognition accuracy by channel concatenation were found only marginal. In [P6] slight gains were again observed, but the difference to monaural features was still minuscule. The foremost reason is that in room recordings the channel separation in the magnitude domain is not very high to begin with, thus magnitude-only NMF does not benefit much from observed differences. Further issues arise from the low spectral resolution of mel features and the additive model, which may combine events across channels arbitrarily. All in all, channel concatenation is not very accurate for modelling stereo data. Even more crucially, it fails to take into account any phase information, which has a major role in dedicated spatial algorithms. Alternative formulations have been proposed for magnitude-only NMF [127] and phase-sensitive complex valued NMF [123, 148, 149], also with a software implementation available [125]. As the prevalence of multi-channel inputs in ASR grows, there is an increasing incentive to combine these methods with other potential directions in NMF frameworks.

3.1.6 Advanced Features and Future Directions

Although already the current multi-frame mel spectrogram features have proven viable in separation and enhancement, there is definitely room for further improvement. Regarding the spectral representation itself, higher resolutions and alternative filter banks such as gammatones or auditory-based features [101, 109, 155, 162] should be considered. Also spectral transforms similar to DCT or other discriminatory steps might help in classification of spectrally close phones, which currently limit the performance of sparse classification [P6]. Any feature space based on fixed band weighting is prone to channel mismatch errors, hence compensation methods have been proposed for adaptive band reweighting [43]. The aforementioned spatial methods and robustness against reverberation are desirable for real-world environments. For modelling of dynamics, semi-negative factorisation [22] may eventually become necessary in order to handle negative transient weights properly. Finally, alternatives to explicit spectro-temporal models, e.g. shorter units with HMMs or delta-driven features, should be studied further to alleviate the exponentially increasing complexity of long-context atoms. Ultimately, feature representations must be developed in conjugation with model acquisition and factorisation methods, which are discussed in the next sections.

3.2 Speech Models

Separation of speech from mixtures in a spectrogram domain relies on the speech model, which is also the only model we can generally assume being available in every scenario. Although there are several delicate factors such as individual voice, stress and emotion affecting the exact spectro-temporal shape of speech patterns, they still lie on a low-dimensional manifold defined by the vocal tract and spoken languages, unlike noise events produced by a plethora of greatly differing sound sources and phenomena. Therefore we can always derive at least an approximate model of speech for any task. Nevertheless, the complex and non-stationary nature of human speech makes the task quite challenging compared to e.g. modelling musical instruments, whose spectro-temporal trajectories are more consistent. In this section, an overview is given to various speech model acquisition methods employed in NMF-based separation and enhancement.

3.2.1 Characteristics of Speech

At any given moment, the spectrum of speech can be approximated as a product of two components. The first is *excitation* produced by air flowing from the lungs, either voiced by resonating vocal folds, or unvoiced if air is flowing freely. Voiced excitation can be modelled as an overtone-rich sequence of harmonics characterised by the *fundamental frequency*, typically ranging from about 85 to 180 Hz. Unvoiced excitation has no distinct harmonic structure, thus it can be approximated as white or coloured noise. The second part is a *filter envelope* formed with the vocal tract, producing the characteristic spectral shapes of vowels and consonants. The envelope is relatively smooth a function, often approximated with a low order all-poles filter. One, two or three spectral peaks known as *formants* can usually be distinguished and suffice for recognising different phones.

In the temporal direction, speech is a continuously changing process, only approximately stationary within single phones, which on average last less than 100 ms including transitions [72]. Typically another phone follows much sooner, drastically changing the spectral shape of produced speech. The dynamics of overall speech energy are roughly defined by syllable-length units with most of the necessary information found within 1–7 Hz modulation frequency range, and up to 12 Hz still helpful for comprehension [27]. In ASR, 100 Hz sampling rate is commonly used for computing the spectrum to ensure sufficient stationarity, and to locate areas of rapid change such as stop consonants.

In natural languages there are recurring units such as syllables, words and complete phrases spanning hundreds of milliseconds with their frequency of appearance decreasing over unit length [140]. However, even in lexically identical units there will be variation due to individual voice profile, intonation, speed, stress, emotion, coarticulation, and natural fluctuations. For any single speaker the variation will be smaller, dictated by physical and learnt attributes of speech production. These individual factors and idiosyncrasies facilitate speaker identification and separation of a single voice from multi-talker mixtures, assuming that an appropriate personal voice model can be acquired. The reduced pattern space also improves ASR results in comparison to speaker-independent recognition, where e.g. the fundamental frequency, tempo and dialect introduce higher variation to speech at all levels. Where speaker-dependent models are not viable or desired, personal traits may be mitigated by using a lower spectral resolution, vocal tract length normalisation (VTLN), or time warping algorithms.

3.2.2 Statistical Frame Models

A widely used representation for speech is statistical modelling of different phones on a frame level. For example, in the default settings of HTK software [213], *mel-frequency cepstral coefficients* (MFCCs) are extracted by computing log-compressed energies from 26 mel bands in 25 ms frames with a 10 ms shift. These are decorrelated with discrete cosine transform (DCT), whereafter the first 12 coefficients are retained, thus modelling the approximate vocal tract envelope shape while discarding a lot of information concerning e.g. the fundamental frequency. Delta and double-delta features are conventionally concatenated to the static spectra to capture temporal dynamics. Finally, phones or sub-word units are modelled statistically as *Gaussian mixture model* (GMM) distributions.

The statistical model is not directly applicable to NMF, where a basis atom is a single point in the feature space instead of a distribution. Also, the log-compression and cosine transform steps are generally skipped for retaining approximate additivity and strict non-negativity. Therefore the fundamental building block of NMF modelling is the original frame spectrum or coefficients from a filter bank as described in Section 3.1, and statistical models are replaced by approximation with summed or nearest spectral vectors.

3.2.3 NMF Learning of Speech

It is possible to learn speech models by applying unsupervised NMF to a corpus of training speech. Low-rank factorisation is used to capture a compressed model of speech spectra. Both single-frame [153] and convolutive multi-frame window [122, 158] algorithms have been proposed for the task. The main benefit of the approach is finding a representative model for any input data with no prior information, annotation or assumptions other than defining the basis dimensions and cost functions. A notable downside is that the additive NMF model may separate units of speech both spectrally and temporally into multiple components. For example, a phone that is characterised by a particular structure of formants may become split into several single-formant atoms, which also activate during other speech or noise events. Consequently the separation and classification capability of such partial atoms is lower than for atoms which model their corresponding speech patterns as a whole.

Models may be learnt in an unsupervised manner with the help of sparsity criteria e.g. for two-speaker separation like was shown in [122], but the resulting units were described only ‘phone-like’ in their appearance. Therefore a higher level of supervision in learning is recommendable for sparse classification or noise robust speech processing to ensure modelling of characteristic large-scale patterns. Already in [153] some supervision was brought into the algorithm by segmenting the training corpus into individual phonemes and learning a separate basis for each. Phoneme-dependent bases were also used in [135]. In [203, 208], each word in the corpus was modelled separately with convolutive single-component learning. In this approach, the resulting atoms could be considered similar to word templates with relatively little actual learning taking place. Because there are frequently arising issues in wholly unsupervised NMF learning of long-term patterns, in this work the preferred methods for speech modelling are using directly sampled exemplars or generalised templates, which contain the complete spectro-temporal profile of their originating speech pattern.

3.2.4 Speech Exemplars and Templates

As justified in Section 3.1.3 and repeatedly observed in experiments [P1, 44, 158, 208], increasing the temporal context of speech atoms is highly beneficial for detecting spectro-temporal speech patterns from complex mixtures. For separation itself, longer context is generally better, and in related work longest matching segment search has been used for enhancement [16, 82]. In NMF, restricting factors arise from increased complexity and uniqueness of long units, which consequently raise computational costs and training data requirements. For small vocabulary corpora, 200–300 ms units have been found manageable [P1, P6, 44]. In other tasks and implementations, shorter or longer units may be preferable as a trade-off between data, cost and quality factors.

In many ASR variants exploiting long context, *exemplars* or *templates* are used to model speech. In this work, exemplars refer to $B \times T$ spectrogram segments sampled directly from external or previously observed data, while templates are their generalised counterparts that still correspond to specific acoustic patterns but no longer to single instances. However, in some other studies no such distinction is made, and the terms are used interchangeably [21, 60]. Not all exemplar or template systems employ NMF, which is apparent already from the fact that the general approach has been used in ASR for over 50 years [18, 66, 97, 134]. Instead, nearest or k-nearest neighbours algorithms have been used for template matching. In many cases, dynamic time warping (DTW) is also incorporated to the model [17, 21, 60]. In NMF-based methods, the focus is on additivity and sparse approximation, whereas temporal flexibility is achieved with alternative means presented in Sections 3.2.7 and 4.1.

The proposed framework originally used speech exemplars sampled randomly from training data [44]. For the equally distributed spoken digit vocabulary of the AURORA-2 corpus [68], random selection of exemplars was found equal or even superior to early attempts of supervised selection. Considerably large a basis, typically 4000 exemplars [44] or even more [40, 48] was used to model the 11-word vocabulary.

In [P2], the exemplar acquisition process was changed slightly. In the 1st CHiME corpus [3] based on GRID speech [14], word frequencies are not uniform between classes, thus random sampling would allocate disproportionately many exemplars on certain parts of the 51-word vocabulary. The proposed algorithm first extracts a dense basis by pseudorandomly stepping through the training data, and then reduces the basis by equalising exemplar counts between words as much as possible. In this method, cross-word exemplars spanning over word borders were allowed and frequent. The framework employing this exemplar model in an optimised feature space still produced the best NMF-based enhancement results on 1st CHiME data in 2013 [87].

A downside of exemplar modelling is that atoms chosen from individual instances of speech do not generalise well to differing pronunciations, thus allocating multiple exemplars for any given pattern is beneficial or even required [48]. Better genericity can be achieved by incorporating concepts of statistical and template modelling to basis acquisition.

In [P6], 1st CHiME speech was alternatively modelled with *templates*, each representing one state of the word-based back-end model, including its neighbouring context. All word spectrograms in the training set containing the sub-word state to be modelled were placed in a $B \times T$ window with the target state maximally centred, whereafter a bin-wise median was taken to form the final $B \times T$ template. An illustration of the process is shown in [P6]. Instead of explicit modelling of instances with a large amount of exemplars, the templates represent approximate spectro-temporal profiles of corresponding speech patterns.

Albeit not as accurate as the large 5000-exemplar bases, the newly acquired bases of 250 templates were 20 times smaller, and only reduced the average recognition accuracy from 86.9% to 85.2% using feature enhancement for a robust back-end. Considering that the average accuracy for unenhanced speech was 74.7%, the quality loss was relatively small compared to the reduction in basis size and consequently computational costs. The 20-fold reduction was simultaneously applied to the noise model too, hence the loss from template modelling of speech alone should be even smaller. Larger losses were observed in sparse classification, though, likely because the template atoms still correspond reasonably well to speech in general, but lose some of their classification accuracy due to spectro-temporal blurring. Therefore they are better suited for enhancement than direct classification. It should also be noted that in enhancement the statistical back-end model that performs state evaluation still retains its full size and complexity, whereas in SC the basis size reduction applies to all components of spectral modelling and classification, hence greater quality losses can be expected in final recognition results.

3.2.5 Segmentation Algorithms for Large Vocabulary

The previously discussed exemplar and template systems [P6, 44] were evaluated on small vocabulary AURORA-2 [68] and 1st CHiME [3] corpora for simplicity in modelling and back-end recognition. Due to their small 11 and 51 word vocabularies, respectively, it is feasible to model each word, their parts, and even all possible transitions between words using long atoms. Conversely, in real-world applications we are ultimately interested in large vocabulary continuous speech recognition (LVCSR), where the same algorithms are not as readily applicable because of the exponentially increasing number of speech patterns.

To illustrate the complexity, already in the baseline back-end of the 2nd CHiME corpus [184] based on medium vocabulary WSJ0 speech, there are:

- 5000 words
- 39 monophones and 2 pauses
- 65561 theoretically possible triphones
- 8784 HMM triphone states after tying
- 4507 tied triphone states found in the training data

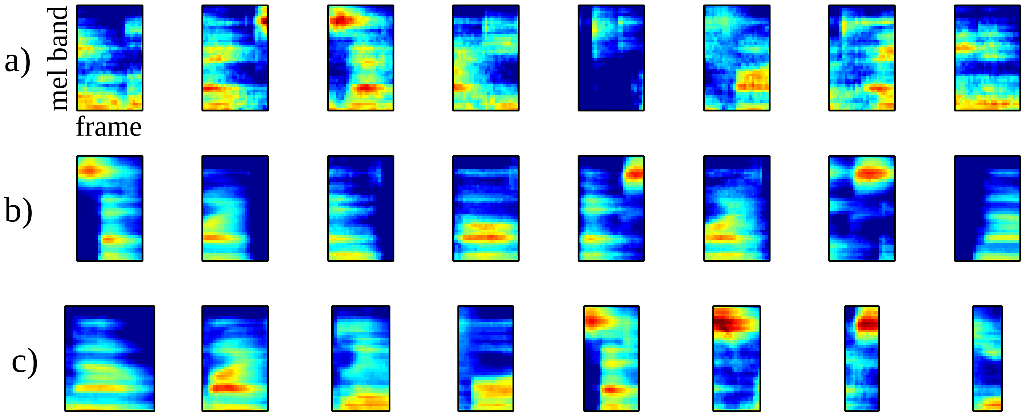
Regardless of considerable reduction of triphone models via tying of phonetically similar forward and backward transitions, the number of different patterns already in a very short context of one phone and its immediate neighbourhood is high for NMF algorithms. For robustness against noise, it would be preferable to use at least syllable-length context or even more. Clearly even rudimentary modelling of each word with a few variants for different speakers would produce an enormous speech basis, which would not be viable for practical purposes. For truly large vocabulary and casual speech the outlook would be even worse.

To find an acceptable compromise between long context and manageable basis size, in [P7] and [P8] a *variable-length* segmentation and clustering algorithm was proposed. Segmentation algorithms in general attempt to find lexically or phonetically meaningful units such as words [1, 121, 138], sub-word units, or phones [72, 89, 129, 131, 151] from continuous speech. For robust ASR, the primary goal of automatic segmentation is to find maximally long units while keeping the basis size practically manageable.

In the proposed algorithm, recurring units are extracted in a decreasing order of length, starting from the maximum atom length we want to use in factorisation. Features are extracted from a training data set, whereafter the algorithm looks for matching frame sequences of a chosen length. A set of matching sequences forms a *cluster*, which is turned into a speech template by averaging the segment spectrograms. When no more clusters of defined size can be found, the window size is decreased by one, and the search is resumed. Consequently a set of variable-length templates is extracted down to a defined minimum length, or until a sufficient percentage of the training data set is covered. The resulting basis will model frequent patterns with maximally long units, typically whole words, while more variable segments are split into syllables or other sub-word units. A variable-length factorisation algorithm is then used to find a convolutive combination of atoms that model the target speech.

To measure the similarity of frames and segments, two data sources are used; mel-spectral features with normalisation and augmentation with deltas, and state sequences from utterance transcriptions and forced alignment. A weighted sum of the two measures is also used. The methods were compared using the small

Figure 3.1: Speech atoms from different acquisition algorithms: a) exemplar sampling, b) fixed-length template modelling, c) variable-length template modelling.



1st CHiME vocabulary, all of them achieving reduction in model size compared to the earlier fixed-length template model, while enhancement and classification quality remained the same or even improved [P7]. For medium vocabulary, small speaker-dependent bases were learnt and used in [P8] for speaker approximation and modelling with a combination of multiple bases. Initial results suggest that variable-length learning is indeed feasible for LVCSR using NMF, although further performance analysis of various novel system components is still needed.

The output of proposed exemplar and template modelling methods is compared in Figure 3.1. Exemplars shown on the first row capture very specific spectro-temporal instances of speech. Averaged template models on the second row display smoother features, which generalise better but are less accurate for classification. On the third row, variable-length templates are shown.

3.2.6 Atom Labelling

For sparse classification, speech atoms are given *label matrices*, which denote the state content of atoms and thus facilitate decoding estimated state likelihoods over target utterances. For large exemplar bases, where the atoms explicitly model their originating spectrograms, sufficiently accurate labels can be obtained simply by storing the corresponding state sequences of the source utterances [P6, 44]. For compact template bases and other indirect basis acquisition methods, strict similarity to originating utterances is lost, hence better labelling has been achieved by employing learning algorithms [70, 106]. These are discussed in Chapter 5.

3.2.7 Alternative Speech Models for NMF

While in this thesis' work exemplar and template spectrogram bases are used to model speech, other variants have also been proposed for NMF systems. In works by Van hamme et al., histograms of acoustic co-occurrences (HACs) are used to compress vector-quantised frame features of variable-length words into single histogram vectors [25, 178]. In [37], single-word utterances are time-warped with interpolation to fixed window length. In [212], multiple factorisations are conducted with speech bases of different atom lengths, whereafter an optimal path across all factorisation results is calculated with dynamic programming. In [7], a generic speech basis is initialised for convolutive modelling and permitted to adapt to a new speaker with a penalty term on its divergence from the original model. Finally, in HMM-based variants, some or all of atoms' rigid temporal context is replaced with transition probabilities, potentially making the system more flexible regarding slight variations in the pace of pronunciation [46, 53, 115, 119].

3.3 Noise Models

3.3.1 About Noise and Its Handling

Speech models alone suffice for separation of concurrent speakers, but in everyday ASR there are practically always noise sources present both in single- and multi-talker scenarios. In reality, very few general assumptions can be made about noise. Even restricting the definition to additive noise sources — omitting reverberation, the acoustic environment, and channel errors — still leaves us with almost an endless range of noise types. They may be strongly localised spectrally (e.g. static tones) or temporally (impact events), wideband (wind, traffic etc.), or complex patterns all the way to competing speech and babble, which obviously appear very similar to target speech. Noise levels can also vary greatly and change rapidly. This enormous variety makes robust ASR a difficult problem with no straightforward solutions.

Recalling the rough categorisation given in Section 1.1.4, robust algorithms can be split into three groups. First, we can try to make the system overall more sensitive to speech and robust against deviations without making assumptions on the exact noise types. For example, auditory features [101, 109, 162] and normalisation methods [181] fall into this category. However, regardless of representation, high levels of noise will introduce increasing variance and unreliability to the features, making the recognition fail due to model mismatch. Therefore in difficult noise conditions it becomes necessary to model noise explicitly in order to enhance the input signal, or to adjust the back-end models accordingly [99, 193]. In this section, noise modelling is discussed from the viewpoint of NMF algorithms.

3.3.2 NMF Models for Noise

The NMF framework presented in this work is in particular a viable tool for explicit noise modelling, thanks to its ability to represent an unlimited amount of overlapping sound sources and audio event types, including highly non-stationary noise [99]. Nevertheless, its performance is strongly dependent on the accuracy of noise models [44].

Many of the methods proposed for modelling speech in Section 3.2 are more or less directly applicable to noise as well. A few notable differences arise from certain characteristics of noise and availability of data. First, the greater diversity of noise events requires either a comprehensive noise dictionary or strong adaptivity to the current environment. Second, noise sources are regularly overlapping with each other, unlike target speech which can generally be assumed a single-source process. Third, the amount of available and matching noise training material may range from plentiful to inexistent. Whereas for speech at least an approximately matching model can be taken for granted, in noise modelling there is a pressing need for an accurate model, yet fewer guarantees of conveniently available and suitable training data.

3.3.3 Exemplar Sampling

The first versions of the proposed framework were strictly supervised, utilising exemplar bases sampled form separate training data [42, 44]. In the AURORA-2 corpus employed, noise training material is available for types corresponding to the test set ‘A’, comprising four environments — subway, car, babble and exhibition hall. Conversely, the environments in test set ‘B’ — restaurant, street, airport, train station — are new, that is, not present in training material [68]. Already in this limited test scenario the significance of model mismatch is clearly visible. The experiments using noise exemplars sampled randomly from the training corpus strongly favoured test set ‘A’, where the noise types match [P1, 42, 44]. The increased temporal context that ideally improves separation of matching data may correspondingly turn detrimental, when the highly specialised atoms no longer model noise patterns encountered in new conditions.

In practice, only in a limited set of applications it is realistic to assume that a matching noise model can be obtained and fixed beforehand. Also, trying to cover maximally many types of noise with a large basis has its drawbacks. Apart from obvious costs of increased memory use and computational complexity, carrying too many noise atoms in factorisation increases the risk of occasional confusion with speech due to partial overlap in spectro-temporal patterns. Ideally, our noise model should cover the actually observed events but nothing else for maximal efficiency and a minimal chance of confusion.

Exemplar sampling can be improved, data permitting, by exploiting the context of target utterances. Typically speech is not throughout continuous, not even in so-called LVCSR. During moments of speech inactivity it is possible to observe the background noise instead, and to update the noise model. An early version of the approach was proposed in [41], utilising the very first and last frames of AURORA-2 target utterances under an assumption that in those frames speech is not present at all or its energy is low. In later work conducted on the 1st CHiME corpus, neighbouring context has been extensively exploited, facilitated by the data set where target utterances are embedded in long noisy sessions at irregular intervals. Already in the first challenge system a noise basis was extracted for each utterance individually by sampling the context forward and backward with a pseudorandom step size [P2]. In [P6] and [35], the same main algorithm was used for the best performing systems. However, in [P6] and later [P7], sparser sampling and less context was used for more compact models, reducing the computational costs by 20 times. Whereas all these experiments exploited perfect knowledge of utterance positions and sampled strictly noise-only segments, in [P4] an alternative system based on *voice activity detection* (VAD) was proposed for locating and sampling background noise, potentially making the overall framework more viable for online and real-time applications. The method is presented in more detail in Section 5.5.2.

3.3.4 Unsupervised Basis Learning

Despite its flexibility and accuracy in a favourable case, exemplar sampling is not without its problems. Especially random sampling of large segments is prone to capturing redundant, insignificant or low-energy events. Partial alleviations were proposed in [P4], where energy thresholding was used to skip near-silent segments in sampling. Continuous basis updating with pruning of least activated atoms ensured that no unused noise features remained in the basis. Still, one further problem arises from the fact that multiple noise events may overlap, and an exemplar sampled from one co-occurrence will not match another combination of events, nor the single-source components alone. Whereas in speech modelling the single-source nature of speech was posed as an argument against NMF learning, in noise modelling the opposite may be justified instead.

Whenever a noise-only segment is located in training data or in the local context, NMF learning can be employed instead of exemplar sampling to find characteristic noise patterns from the segment. As in other unsupervised learning tasks, a low number of components and optional sparsity constraints act as restricting factors to prevent overlearning and fragmentation of patterns. In [205, 206], single-frame NMF with temporal regularisation was used for learning noise bases from training data. In [185], noise bases were learnt with NMD algorithms from separate training data and the context. In [203], NMD learning was applied to general training data, while the local context was modelled with exemplars. In [P8], a part of the noise model was learnt from training data, again using NMD.

Because NMF algorithms become computationally burdensome for large corpora, typically the training data is segmented and sometimes reduced before applying NMD [P8, 203, 208]. In [202], a two-stage learning process was proposed. In this variant, manageably sized blocks are factored first, whereafter the learnt bases are concatenated and factored again to remove redundancy across blocks. An online learning algorithm has also been proposed in [185, 197] to reduce the complexity of large learning tasks.

3.3.5 Semi-Supervised Factorisation

As pointed out in this chapter, the general features of speech are reasonably predictable, thus at least a rudimentary model for it can be assumed to exist. Meanwhile, the variance in noise patterns is far greater, and there is no universal guarantee of a matching training corpus or even local context. Even in otherwise favourable cases, there is always a chance that completely new noise events are only observed overlapping the target speech. These issues motivate *semi-supervised factorisation*, which has been proposed by several authors [115, 165, 202]. The fundamental idea is deriving a fixed speech basis in advance, but adapting some or all of the noise atoms on the fly.

Using the simple NMF model given in Equation (2.3), let us divide the basis and activations into speech and noise halves, denoted by s and n superscripts, respectively. The model becomes

$$\mathbf{Y} \approx \begin{bmatrix}\mathbf{A}^{\mathrm{s}} & \mathbf{A}^{\mathrm{n}}\end{bmatrix} \begin{bmatrix}\mathbf{X}^{\mathrm{s}} \\ \mathbf{X}^{\mathrm{n}}\end{bmatrix}, \tag{3.1}$$

where only $\mathbf{A}^{\mathrm{s}}$ is fixed while $\mathbf{A}^{\mathrm{n}}$ and both activation matrices $\mathbf{X}$ are updated during minimisation of the cost function. The noise basis can be split further into fixed and adaptive parts to combine known and online-adapted noise models.

Similarly to speech and offline noise basis learning, semi-supervised factorisation algorithms have been proposed employing single-frame atoms [118, 200, 202], HMMs with regularisation [115], multiple speech bases with an adaptive noise model [165], and a convolutive model [P6, 7]. In [P4], a combination of context-sampled and wholly adaptive noise atoms was used for modelling evolving noise conditions in continuous inputs. In [P8], noise atoms were learnt using NMD over an extended span containing both pure noise and mixed-content segments, and also in combination with a fixed noise basis.

However, online adaptation of noise models from mixtures bears a risk that the learnt atoms also capture speech features. As an extreme example, given a matrix

$\mathbf{Y}$ containing noisy speech, the model described in Equation (3.1) may lead to a solution of the form

$$\mathbf{Y} \approx \begin{bmatrix} \mathbf{A}^{\mathrm{s}} & \hat{\mathbf{A}}^{\mathrm{n}} \end{bmatrix} \begin{bmatrix} \mathbf{0} \\ \hat{\mathbf{X}}^{\mathrm{n}} \end{bmatrix}, \tag{3.2}$$

where $\hat{\mathbf{A}}^{\mathrm{n}}$ contains the features of both $\mathbf{A}^{\mathrm{s}}$ and $\mathbf{A}^{\mathrm{n}}$, and $\hat{\mathbf{X}}^{\mathrm{n}}$ changes accordingly. That is, the adapted noise basis models not only actual noise but also the overlapping speech features. Crucially, in a sparse model this solution may have a lower cost than the correct speech-noise activation pattern that employs the actual speech atoms. Even if the loss of speech features is not complete, overadaptation will harm enhancement and classification, because the characteristic spectral profiles of phones become corrupted.

In [P6], semi-supervised factorisation produced uniform improvements in recognition of 1st CHiME speech using both feature enhancement and sparse classification based methods. Conversely, in the experiments on 2nd CHiME Challenge data derived from more complex WSJ speech, greater uncertainty was observed even in semi-supervised factorisation exploiting the noise context. Incorporating a separate fixed noise basis was found to improve the results significantly [P8]. There is ongoing research trying to address the trade-off between adaptivity and overmodelling of noise in the semi-supervised approach. Defining priors on the noise models and tuning of the sparsity parameters are some of the tools to be considered for reducing overadaptation.

3.3.6 Combinations and Further Alternatives

The NMF approach as a whole is quite flexible concerning the levels of supervision and their combinations even within a single factorisation task. It is entirely feasible to use fixed exemplars and learnt atoms from offline acquisition, while also adapting a part of the basis in a semi-supervised manner. Examples of these combinations can be found in [P4, P8, 203]. In [41], artificial noise atoms consisting of single-band features in multi-frame windows were used to capture stationary noise features. In [77], a traditional stationary noise model was integrated into an NMF framework. There are also preliminary results on combining additive noise modelling with compensation of channel mismatch [43].

The whole set of tools comprising exemplars, learnt atoms, synthetic models, online adaptation, priors, deconvolution, channel compensation, and spatial methods provides plenty of options and potential paths for addressing the challenges of real world noise. Arguably the greatest current open issue is finding an efficient combination of algorithms and parameters, which would be generally applicable to the great variety of everyday conditions and recognition tasks.

Chapter 4

Factorisation Algorithms

In the very centre of the systems presented in this thesis is the factorisation process. As its inputs it takes the observation spectrogram, complete or partial source models, and parameters for factorisation. The output comprises primarily activation weights, which are used for classification, spectrogram estimation, and enhancement. Sometimes learnt basis atoms are the main or supplementary output.

The role of factorisation depends on how it is viewed. As a stage in the overall task it is undoubtedly integral. On the other hand, it could be argued that factorisation itself is a trivial computational step, which produces results already defined by the models and parameters. However, the lack of closed form solutions for typical NMF problems means that solving large factorisation tasks is far from trivial regarding practical resources and algorithm choices.

Instead of repeating mathematical or numeric implementations of NMF algorithms, which are better detailed in literature [12] and briefly discussed in Chapter 6, this chapter describes different variants of NMF, especially in the context of audio spectrograms and multi-source separation tasks. It is shown how the spectro-temporal behaviour of speech and noise favours certain algorithms and modelling methods. Later in the chapter, algorithms for handling multiple source models and their grouping are discussed.

4.1 Temporal Continuity

Whereas many NMF applications concern factorisation of data sets consisting of multiple independent data points, a defining feature of spectrogram factorisation is that common sound sources bear strong temporal connectivity, i.e. an audio event forms a spectro-temporal pattern over a time span, which may range from fractions of a second upward with no upper limit. In a frame-based spectrogram representation, an event typically comprises multiple frames with more or less evolving behaviour over time. Assuming that a sequence of spectral feature vectors $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ has already been observed, they set a strong prior for $\mathbf{y}_{n+1}$, defined by the characteristic trajectories of sources. Similarly the content of a corrupted vector $\mathbf{y}_i$ in the middle of a sequence can be predicted from its neighbours, even if the same problem would be unsolvable for isolated vectors.

In multi-source scenarios, spectro-temporal events will regularly overlap each other, forming a continuous mixture of patterns, each following its own trajectory. We refer to this behaviour as *temporal continuity* for short. NMF algorithms favouring general continuity in audio spectrograms were proposed already in 2003 [187, 189] with quality improvements observed in music applications. However, speech and dynamic noise sources do not remain as stationary over time as instruments, thus another set of algorithms is desirable for robust speech processing.

4.1.1 Single-Frame Separation

The simplest approach to spectral separation is deriving the source models and observation vectors from single frames, treating them as unordered sets with no explicit continuity. In other words, a random permutation could be performed on the columns of observation spectrograms with no other effect than the same permutation taking place in activations. This is also how early speech separation examples [153] and other studies in audio separation [62, 159, 195] were presented. Obviously such an algorithm does not address continuity at all, regularly leading into ill-posed problems when similar-appearing sources must be separated from individual frames without context. Slight improvements can be achieved with the previously mentioned continuity criteria [187, 189] treating the difference in activation weights between consecutive frames as another objective to be minimised, by inter-frame regularisation [205], forward-backward smoothing [114], or by including delta features in the frame vectors as described in Section 3.1.4. Nevertheless, audio events and temporal correlations in speech span significantly longer durations than what can be modelled with frame vectors even with augmentation [140]. More generally, the non-stationary nature of speech and many noise sources eventually favours more explicit modelling of long-term patterns.

4.1.2 Multi-Frame Windows

A common method to incorporate more context into NMF models is to extend the models and observation periods from single frames to multi-frame *windows* as described in Section 3.1.3. In *sliding window factorisation*, atoms and observation windows comprise T consecutive frames in the spectral feature space of B bands. These are reshaped into $TB \times 1$ vectors, whereafter such multi-frame feature vectors can be treated equivalently to their single-frame counterparts, only with a longer context explicitly captured in them [44]. In the factorisation stage, overlapping windows are extracted from the observation spectrogram, hence the ‘sliding window’ nomenclature. The overlap produces several independent estimates for frames appearing in multiple windows, which are typically averaged in later processing. The sliding window method was proposed in [47], and used repeatedly in experiments on AURORA-2 data e.g. in [44, 48] and also for 1st CHiME Challenge data in [P2, P3, P6].

From a computational point of view, especially the supervised version of the method is easy to implement as it effectively falls under the basic feature vector model of NMF. Regarding computational costs, data matrix sizes and factorisation complexity increase approximately linearly to window length T, assuming that the sampling interval remains the same, i.e. 1-frame step is used in extracting the observation windows. Whereas in single-frame factorisation of a length T_{utt} utterance the dimensions of matrices in equation $\mathbf{Y} \approx \mathbf{AX}$ are $B \times T_{\text{utt}}$, $B \times L$ and $L \times T_{\text{utt}}$, respectively, in sliding window factorisation they become $TB \times W$, $TB \times L$ and $L \times W$, where $W = T_{\text{utt}} - T + 1$ is the number of factorisation windows. Observation vectors can still be factored independently with parallel or distributed computing. Temporal connectivity modelled by the method takes place within atoms and windows, but not explicitly for any longer periods in the observation.

All in all, the sliding window approach is a simple and reliable way to extend the temporal context of spectrogram factorisation. However, one of its foremost problems is its strict dependence on temporal alignment of events. Let us illustrate the matter with an example, where a basis atom perfectly models the events found in observation frames $1 \ldots T$. In the next window, covering observation frames $2 \ldots T + 1$, the features have shifted by one column. The previously used atom no longer matches accurately, despite that most of the patterns in the observation window remain the same, only with a 1-frame shift. Around a window where an atom matches the observation, there are $T - 1$ partially overlapping windows to both directions, which would require shifted variants of the same atom. Thereby up to $1 + 2(T - 1) = 2T - 1$ different atoms, each corresponding to a different temporal alignment, would be needed for perfect representation. If the basis falls short of these requirements, estimation will be inaccurate in observation windows where a correctly aligned atom cannot be found. Although averaging over multiple overlapping window estimates will reduce the severity of the issue, the temporal model is fundamentally inflexible. The problem carries over to semi- and unsupervised modelling, where atoms should be learnt online. NMF learning generally relies on finding a low-rank model, which contradicts with the sliding window model’s preference for multiple shifted variants of each event. For these reasons, the method appears best suited for supervised factorisation with an overcomplete basis, while other scenarios may favour the convolutive version.

4.1.3 Convolutive Modelling

Convolutive non-negative matrix factorisation (CNMF), in this work referred to as non-negative matrix deconvolution (NMD), is an alternative model for representing observations consisting of temporally connected events [152, 157, 188]. In fact, it was proposed for speech applications even earlier than the sliding window method with examples on two-speaker separation and speech denoising [122, 158]. Complete evaluations on noise robust ASR via enhancement and sparse classification were presented in 2011 [P1, 185, 203], whereafter the method has been widely used and extended in speech applications [P5, P8, 34, 186, 208, 210].

In many respects, NMD bears high similarity to sliding window processing. In both approaches the basis consists of length T atoms, which are activated over time to model the observation spectrogram with extended context. The crucial difference arises from interpretation of activations concerning the spectrogram estimation. In sliding window NMF, each window is considered a separate factorisation task and processed independently of its neighbours even when the windows overlap. In NMD, the estimated observation $\boldsymbol{\Psi}$ is calculated convolutively over all time indices, thus the activations jointly produce a single length T_{utt} estimate, which is the sum of all atom spectrograms emitted by timed activations.

Mathematically, the model can be presented in an intuitive manner as

$$\boldsymbol{\Psi} = \sum_{l=1}^{L} \sum_{w=1}^{W} (x_{l,w} \overset{\rightarrow(w-1)}{\mathbf{A}_l}), \tag{4.1}$$

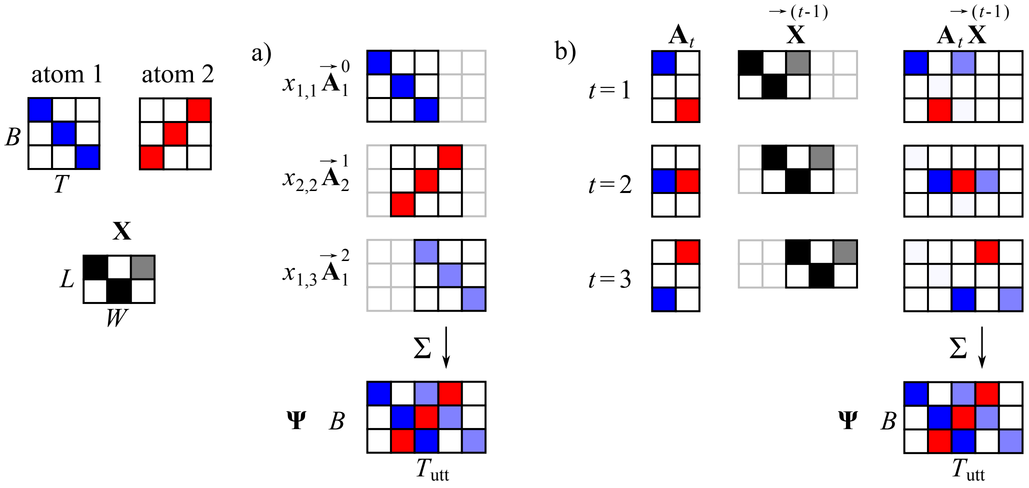
where index $l \in [1, L]$ is used for atoms and w for windows from 1 to $W \leq T_{\mathrm{utt}}$. Scalar $x_{l,w}$ is the corresponding activation weight, $\mathbf{A}_l$ is the $B \times T$ spectrogram of atom l, and operator $\rightarrow$ shifts it right by $w - 1$ columns in a $B \times T_{\mathrm{utt}}$ zero-padded matrix. In other words, the overall estimate is a sum of single-atom estimates, each consisting of a sum of atom spectrograms weighted and time-shifted according to the activation pattern. This approach is illustrated in the first half of Figure 4.1, where two 3×3 atoms and a 2×3 activation matrix produce a 3×5 spectrogram. Three non-zero activations and corresponding shift operations take place.

However, for practical efficiency, the same equation is conventionally given as

$$\boldsymbol{\Psi} = \sum_{t=1}^{T} \mathbf{A}_t \overset{\rightarrow(t-1)}{\mathbf{X}}, \tag{4.2}$$

where t indexes atom frames from 1 to T, $\mathbf{A}_t$ is a $B \times L$ matrix containing t^{th} frame vectors of all atoms, and $\rightarrow$ shifts the $L \times W$ activation matrix right by $t - 1$ columns in a $L \times T_{\mathrm{utt}}$ zero-padded matrix. An illustration for the same data as in the previous example is seen in the second half of Figure 4.1. Three $\mathbf{A}_t$ matrices multiplied by shifted $\mathbf{X}$ form the partial estimates and finally the same

Figure 4.1: The convolutive model and its computation. Two 3×3 atoms (colour-coded for clarity) and a 2×3 activation matrix produce a 3×5 spectrogram in the non-truncating convention. In algorithm a), atoms are shifted and weighted for each non-zero activation. In algorithm b), atom-frame matrices are multiplied by the shifted activation matrix for each t value. Both produce the same output.


atom 1
atom 2
B
T
X
L
W
a)
$x_{1,1}\overset{\rightarrow 0}{\mathbf{A}_1}$
$x_{2,2}\overset{\rightarrow 1}{\mathbf{A}_2}$
$x_{1,3}\overset{\rightarrow 2}{\mathbf{A}_1}$
Σ
Ψ
B
T_{utt}
b)
$\mathbf{A}_t$
$\overset{\rightarrow (t-1)}{\mathbf{X}}$
$\mathbf{A}_t\overset{\rightarrow (t-1)}{\mathbf{X}}$
$t=1$
$t=2$
$t=3$
Σ
Ψ
B
T_{utt}

result as in the first approach. This reformulation is often preferred, because it reduces the computation to a single loop with only one matrix shift per step, and T is typically the shortest dimension for actual basis arrays. Most operations are performed with large matrix multiplications, which are usually well optimised and suitable for parallelisation. Iterative update rules for solving the basis and activation matrices are given for different cost terms in literature [12, 158].

There are two popular choices for the maximum temporal index W for activations. If we set it to $T_{\text{utt}} - T + 1$, all atom spectrograms produced by activations will fit entirely in the estimated observation matrix, similarly to the sliding window convention of only extracting full windows from $\mathbf{Y}$. If it is set to T_{utt}, the last $T - 1$ activations will produce truncated atom spectrograms, but the temporal dimensions of $\mathbf{Y}$ and $\mathbf{X}$ will match. For compatibility with earlier sliding window work and convenient simplifications in computation, this work uses the former convention, although the latter also appears often in other work [12, 158].

In typical cases, the activation pattern produced by NMD will be considerably sparser than the corresponding sliding window activation output. The reason is that a single NMD activation may model a complete length T event in the estimated observation spectrogram, hence the neighbouring activation indices can be empty. In sliding window NMF, overlapping windows are factored independently, and they all contain spectral features which must be modelled separately with activations, thus producing approximately T times more overall activation weights and a smoother activity pattern across consecutive windows. The same behaviour also makes NMD more time invariant. Fewer temporal alignments of events are required in the basis, because the algorithm will find a sparse set of best fitting temporal positions for activations across the whole observation.

Although the fundamental complexity class of solving NMD is equivalent to sliding window factorisation of the same size, its textbook implementations involve looping over at least one dimension of a 3-D basis array, followed by summing with matrix shifts, eventually making it slower in practice. However, optimised versions for parallel computing are being developed, and the complexity gap is narrowing down, especially considering the potentially smaller NMD bases.

In robust ASR, results from comparisons of the two temporal models are still inconclusive. In [180], convolutive modelling was rejected due to its higher computational costs and worse performance in robust pattern learning. In [P1], NMD was found to perform mostly better than sliding window NMF in small vocabulary SC with identical speech and noise bases, although in mismatching noise conditions its performance deteriorated more heavily. In [P2], the sliding window method performed better for the 1st CHiME data using large exemplar bases, thus it was later used in [P3] and [P6]. However, in [P6] NMD was in turn found better for separation with small speech and noise bases. In [51], NMD performed better than single-frame separation with post-smoothing but worse than sliding window factorisation, measured by computational metrics in speech-music separation.

These results suggest that the sliding window approach may be more robust especially in mismatching noise conditions due to its averaging over multiple estimates, when large enough bases can be used to provide sufficiently many variants for different temporal alignments. Conversely, the sparse activation pattern of NMD may produce more misclassifications in mismatched modelling, but its temporal invariance is more efficient for compact modelling of diverse speech and noise patterns. Moreover, NMD facilitates advanced techniques such as unsupervised or semi-supervised learning and variable length modelling. For now, both models should still be studied further, because results of comparisons are not conclusive yet. Ultimately the output quality depends on several factors such as the separation task, basis size, and factorisation parameters.

4.1.4 Variable Unit Length

As pointed out in Sections 3.1.3 and 3.2.5, speech consists of diverse units, whose duration and rate of occurrence vary considerably. Obviously the same also applies to noise. There is thus motivation for modelling variable-length acoustic units, which can also be acquired by segmenting and learning algorithms.

Earlier work aiming at variable-length modelling with NMF or templates includes interpolation to fixed-length arrays [37], using histograms of quantised data points [139, 178], and variable-length templates for latent perceptual mapping [164]. Longest matching segment search has been used for dereverberation [82] and robust ASR [16].

Commonly presented NMF and NMD algorithms only operate on fixed-length units. However, explicit modelling with varying atom length is also possible, albeit only recently applied to speech processing. Sliding window processing effectively requires fixed-length units in any single factorisation task. In [212], multiple factorisation passes are conducted with different atom lengths, and a search algorithm is used to find the best matching lexical sequence across factorisations. The requirement for fixed window and atom length does not apply to NMD, though, because its observation estimate $\mathbf{\Psi}$ is a single array of atom spectrograms activated over time, permitting any combination of atom lengths.

In [196], a variable-length NMD model and preliminary speech separation results were presented. The implementation is relatively straightforward. Starting from the estimation model of Equations (4.1) and (4.2), we introduce variable atom length T_l depending on the atom index l. Slight rearrangement results in formulation

$$\mathbf{\Psi} = \sum_{l=1}^{L} \sum_{t=1}^{T_l} \mathbf{A}_{l,t} \overset{\rightarrow(t-1)}{\mathbf{X}_l} , \tag{4.3}$$

where $\mathbf{A}_{l,t}$ is the t^{th} frame column of atom l, and $\mathbf{X}_l$ is the l^{th} row of the activation matrix. Similarly to fixed-length NMD, activation time indices may be permitted up to atom-dependent $W_l = T_{\text{utt}} - T_l + 1$ or T_{utt}, depending on whether partially fitting activations are permitted. Atom lengths T_l ideally correspond to the durations of modelled events. In an informed case it may be possible to define the distribution beforehand. However, real-world audio material is more likely to call for data-driven methods like incremental search [196] or algorithmic segmentation of a training corpus [P7, P8].

Variable-length speech models were used for small vocabulary in [P7] and for medium vocabulary in [P8]. Preliminary results suggest that more efficient modelling is indeed possible by adjusting the unit length according to the content of speech patterns. Also, longer units can be included in the basis than would be viable in fixed-length modelling. The approach as a whole is still in an early stage. As future research topics, better segmentation algorithms and variable-length noise atoms should be studied for improved modelling and robustness.

4.1.5 Hidden Markov Models

A persistent problem in long-context atom modelling using exemplars, templates, or other acquisition methods is that atoms are rigid units of considerable length, whereas real speech and noise features often vary slightly in their pace and duration. Therefore e.g. an exemplar sampled from one speech instance may differ in its temporal behaviour from another observed instance, even from the same speaker. In common ASR systems, speech is modelled with short-time feature vectors controlled by a hidden Markov model (HMM) for representing the likelihoods of advancing from one linguistic state to another. In a conventional HMM, the time spent in any state is not a constant but an exponentially decaying distribution of probabilities. Several different transitions may also be permitted, making the model more flexible regarding phone durations and acoustic trajectories.

The HMM approach is actually viable for NMF frameworks as well, and it has been demonstrated several times in literature. Initial versions evolved from *factorial scaled HMMs* (FS-HMM) [124] to *non-negative HMM* (N-HMM) and *non-negative factorial HMM* (N-FHMM) [119], which was later extended to its semi-supervised version [118]. Further variants have been studied by Grais and Erdogan [53], Gemmeke et al. [46], and Mohammadiha et al. [115]. An overview to dynamic NMF models is given in [160].

Generally speaking, HMM regularisation tends to improve results over unregularised factorisation by adding temporal connectivity that is missing from the baseline model. However, HMMs are not without their issues either. Crucially, the Markovian model assumes that each transition only depends on the state itself, not its preceding sequence, which is not true e.g. for speech [140]. Expanding the feature vectors with concatenated frames or deltas will introduce more context to the model, but it in turn violates the Markovian assumption further [140]. Nevertheless, more flexible transition models have their appeal for reducing the rigidity and thus large basis sizes of long-context event modelling. HMMs, despite their shortcomings, are a well studied and understood tool for speech modelling, and they should be still considered for NMF and hybrid implementations. They are not in the main focus of the work covered by this thesis, but they have appeared as an experimental addition to the framework in [46].

4.2 Block-Mode Processing

The last part of long-context modelling discussed here is selecting the size of segments that form observation matrices in factorisation. While not an issue in supervised sliding window factorisation, where each window can be extracted and factored individually, the choice becomes highly relevant in the joint model of NMD and especially its semi- or unsupervised variants with basis adaptation.

In many artificial evaluation tasks, the observation is a well defined ‘utterance’, which typically spans up to a few seconds, and has been endpointed to contain only speech. In real-world tasks, we cannot assume that speech is segmented into such convenient pieces. Although several voice activity detection (VAD) algorithms exist for approximating the on- and offset times of speech [108, 144, 215], there is an unpredictable delay in both speech and noise segment processing until we know where the segment ends. Very long observation spectrograms also become inconvenient to handle with NMF algorithms, even more so in NMD, which must reconstruct the whole observation matrix at once.

In [P4], an algorithm is proposed for processing continuous inputs, where utterances appear at unknown intervals over an evolving noise background, thus calling for constant adaptation. The proposed system uses *block-mode processing*, extracting fixed-size blocks (7.5 seconds) from the input. Each new block is factored using multiple speech bases and the most recently updated noise model. An NMF-based VAD algorithm is used to mark input ranges as speech and noise, whereafter speech segments are re-factored more accurately for enhancing the target utterances, while noise segments are used to update the noise model adaptively. The framework introduces a novel combination of NMF tools for multiple purposes including discovery of speech segments, noise adaptation, speaker identification, and robust ASR from continuous inputs.

Block processing guarantees an upper limit on the processing delay and factorisation task size, which are both desirable characteristics for real-world systems. The results show that an efficient noise model can be maintained by updating and pruning the basis from a continuous input, and that enhancement quality close to oracle segmentation and adaptation is achievable with heuristic updating.

4.3 Multiple Bases and Group Constraints

Because there is notable individual variation in natural speech production, including physical attributes of the vocal tract, speaking style, pace, dialect and so on, it is always desirable to use a maximally well-matching speaker model in ASR. If no assumptions or estimates can be made on the active speaker's identity, a generic, speaker-independent model is often used instead at the cost of slightly reduced accuracy. However, in separation and robust ASR tasks, a discriminative noise profile may be crucial for discovering the target speech among other voices in the spectrogram domain. Individual voice models are also required for speaker identification tasks, where a matching identity must be selected or verified.

In exemplar-based systems, atoms are acquired from individual speakers with no statistical modelling. This is not a problem in scenarios where the target speaker's identity is known, and a matching model can be trained beforehand as in [P2, P6]. Unknown, newly introduced, and mixed speakers complicate the matter, though. To make an exemplar-based system speaker-independent, a large set of training speakers can be used to form a comprehensive basis covering several voice profiles. This was the approach used in early work on AURORA-2 [P1, 44]. Obviously a mixed basis with no further constraints will match many speakers and their combinations simultaneously, hence it alone cannot e.g. separate two speakers from each other. The same applies to generic template models acquired by averaging over speakers.

An alternative approach to speaker identification and approximation problems is to use a structured set of multiple speaker-dependent speech bases. In the combined basis, each atom's originating speaker is known. Therefore by comparing the relative activation weights of atoms across bases it is possible to estimate the exact or approximate voice profile of the observed speaker. The method was first used in [P4] for enhancement and recognition of utterances from unknown speakers using the best matching speaker model or their combination, and in [143] for robust speaker identification. Both experiments were conducted on the closed set of 34 GRID/CHiME speakers [3, 14].

In [P5], a *group sparsity constraint* was introduced for favouring solutions, where a sparse set of speaker-dependent bases is active instead of an unrestricted combination. Assuming that the overall basis $\mathbf{A}$ comprises N source-specific groups $\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \ldots, \mathbf{A}^{(N)}$, and activations are correspondingly denoted by N vectors $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(N)}$, a penalty function of form

$$f = \sum_{n=1}^{N} \left\| \mathbf{x}^{(n)} \right\|_2 \tag{4.4}$$

is introduced to overall cost minimisation. The proposed function measures the L_2 norms of activations within single speech bases, and then their sum (L_1 norm) across bases. Consequently, multiple activations from the same speech basis have a smaller cost than the same weight coefficients distributed over multiple bases, and modelling is more likely to happen with atoms from a few sources. Improvements were observed in speaker and speech recognition accuracy [P5]. In [P8], new speakers are approximated using a constrained combination of bases acquired from a disjoint training set by employing the same group sparsity term.

Other cost functions have been proposed in [165] for similar purposes, namely performing speech enhancement using a combined basis with a preference for a narrowed down set of speech models. Group sparsity and clustering methods have also been used in NMF applications in [93, 105, 172]. As the approach is still quite novel, there is a lot of room for refining the cost functions and constraints. From a computational point of view, there is appeal in using a rapidly converging algorithm for multiple bases, because it permits pruning of inactive models and hence smaller basis sizes over further iterations like proposed in [P8].

Chapter 5

Separation and Recognition

The goals of factorisation-based speech processing are varied. In some applications, signals are separated or enhanced for human listeners or for further computational processing of the target signal. In other tasks, sufficient information can be extracted from factorisation weights and atoms without returning to the time domain or spectrogram representation. Sometimes the output is a simple decision like a yes/no answer in speaker verification or voice activity detection. On the other hand, deriving the complete lexical content of continuous speech is a complex task involving comprehensive language models and careful evaluation of numerous hypotheses. This chapter primarily describes the main paths of factorisation-based speech recognition, but alternative purposes and methods are also discussed.

5.1 Source Separation

For any well defined set of concurrent audio sources, an innate task is to separate them into their own feature or signal streams. Examples include multi-speaker separation and audio scene analysis, although similar algorithms apply regardless of the types of sources.

In the presented NMF systems, the process generally starts from spectrogram estimation models such as Equations (2.2) and (4.2). Assuming that the atoms can be assigned to N source-specific bases $\mathbf{A}^{(1)} \ldots \mathbf{A}^{(N)}$ with their corresponding activation weight subsets found in submatrices $\mathbf{X}^{(1)} \ldots \mathbf{X}^{(N)}$, applying the estimation formula to each group individually will produce N spectrogram estimates $\mathbf{\Psi}^{(1)} \ldots \mathbf{\Psi}^{(N)}$, each representing the features belonging to a single source. Although these source-specific estimates are already superficially valid, in practice they are only approximate, displaying inaccurate behaviour caused by inherent basis mismatch, bias from sparsity constraints, and other modelling issues.

The spectro-temporal behaviour of the original mixture is usually better approximated by using a wiener-like filter so that a modified source estimate $\hat{\boldsymbol{\Psi}}^{(n)}$ is computed by distributing the original spectrogram content binwise according to the relative estimate weights,

$$\hat{\boldsymbol{\Psi}}^{(n)} = \frac{\boldsymbol{\Psi}^{(n)}}{\boldsymbol{\Psi}} \otimes \mathbf{Y}, \tag{5.1}$$

where matrix division and multiplication with $\otimes$ are performed binwise. The total estimate $\boldsymbol{\Psi} = \sum_{n=1}^{N} \boldsymbol{\Psi}^{(n)}$, exactly, because starting from the vector model of Equation (2.2),

$$\begin{aligned}
\boldsymbol{\psi} &= \mathbf{A}\mathbf{x} \\
&= \left[\mathbf{A}^{(1)} \ldots \mathbf{A}^{(N)}\right] \begin{bmatrix} \mathbf{x}^{(1)} \\ \vdots \\ \mathbf{x}^{(N)} \end{bmatrix} \\
&= \mathbf{A}^{(1)}\mathbf{x}^{(1)} + \ldots + \mathbf{A}^{(N)}\mathbf{x}^{(N)} \\
&= \boldsymbol{\psi}^{(1)} + \ldots + \boldsymbol{\psi}^{(N)}.
\end{aligned} \tag{5.2}$$

Similar derivation applies to overlapping windows and the convolutive case, which are simply summed or averaged variants of their vector counterparts, thus linear operations. Consequently, the sum of weighted estimates

$$\begin{aligned}
\hat{\boldsymbol{\Psi}} &= \sum_{n=1}^{N} \hat{\boldsymbol{\Psi}}^{(n)} \\
&= \sum_{n=1}^{N} (\frac{\boldsymbol{\Psi}^{(n)}}{\boldsymbol{\Psi}} \otimes \mathbf{Y}) \\
&= \frac{\sum_{n=1}^{N} \boldsymbol{\Psi}^{(n)}}{\boldsymbol{\Psi}} \otimes \mathbf{Y} \\
&= \mathbf{Y},
\end{aligned} \tag{5.3}$$

so the estimates together equal the original spectrogram, which would generally not be the case if the $\boldsymbol{\psi} = \mathbf{A}\mathbf{x}$ vectors or corresponding matrices were used as the separation output.

The described separation takes place in a real-valued magnitude or energy spectrogram space, thus omitting all phase information. For many purposes like computing spectral features for common ASR back-ends, the magnitude-only features already suffice. However, to return the estimates to the time domain, phase information must be reintroduced to the spectrograms. Although a variety of phase estimation methods exist [57, 96, 156, 163], in this work the original phase of the mixture is used for all components. Assuming that the sources are mostly sparse and isolated in the spectrogram domain, plausible results are achieved even with this simple approach, making it a common choice in literature [158, 189]. Signal synthesis from mel magnitudes also involves mapping the mel filter weights back into DFT resolution by a matrix transpose or pseudoinverse [P6], and overlap-addition corresponding to the original feature extraction method.

5.2 Speech Enhancement

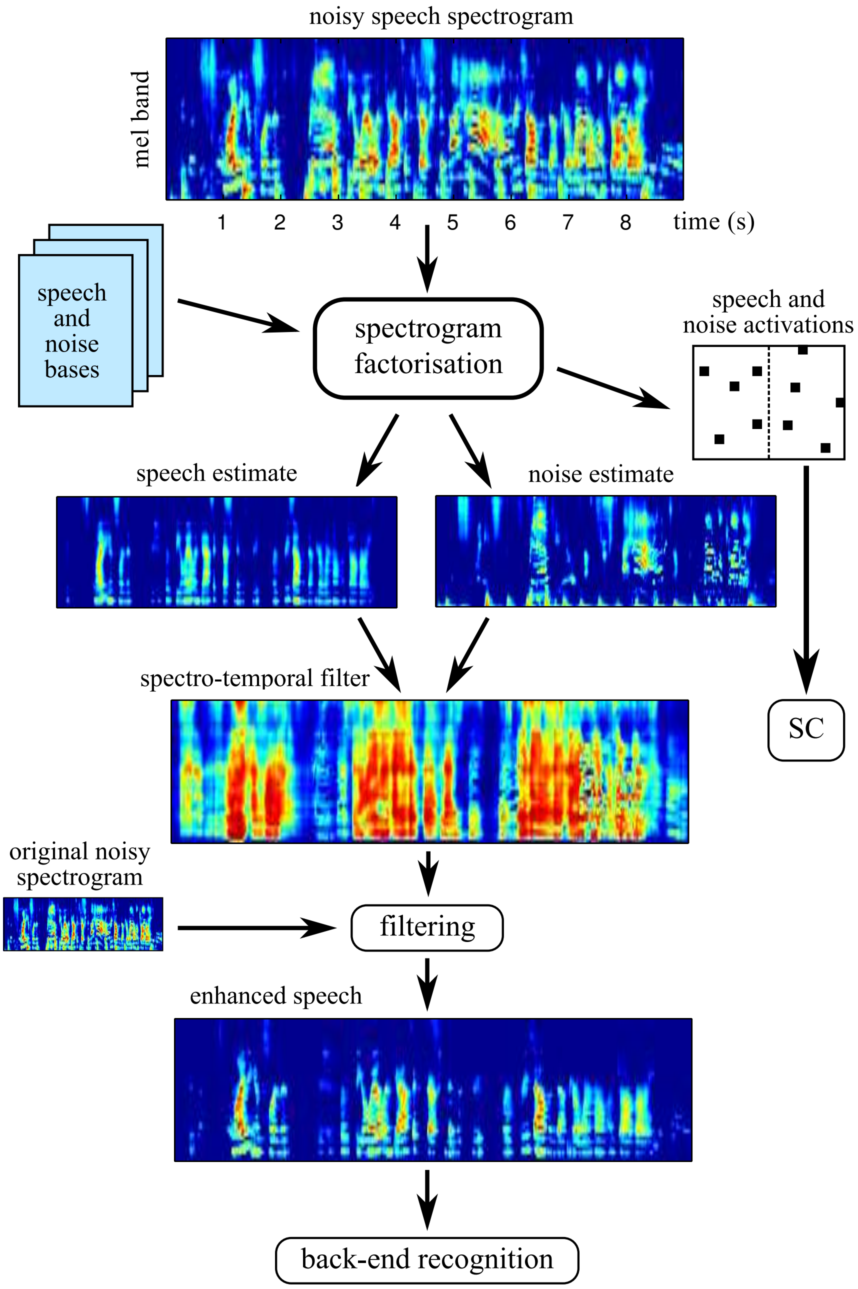
In a common NMF framework, speech enhancement is essentially a special case of generic separation. A speech spectrogram estimate $\mathbf{\Psi}^{\mathrm{s}}$ is computed from atoms $\mathbf{A}^{\mathrm{s}}$ and activations $\mathbf{X}^{\mathrm{s}}$ belonging to the target speaker using equations like (2.2) and (4.2). It is compared to the overall estimate $\mathbf{\Psi}$, producing the bin-wise spectrogram filter weight matrix as $\mathbf{\Psi}^{\mathrm{s}}/\mathbf{\Psi}$, which is then used to filter the original spectrogram.

A schematic diagram of the process with a sample utterance is show in Figure 5.1. Speech and noise spectrogram estimates are computed with NMF. They form the previously described filter matrix, which is finally applied to the original noisy spectrogram to obtain the enhanced speech spectrogram. Optionally, post-processing operations such as spectro-temporal smoothing may be applied for reducing artifacts from the plain filter, thus improving perceived quality or signal behaviour in further processing. Post-processing has been observed to improve output quality measured by computational metrics in speech-music separation [54, 56]. However, none was used in the work included in this thesis.

In a consistently designed spectrogram enhancement and ASR system, there is no need to go all the way back into signal level, because typical back-end features can be derived from the DFT spectrum or even compressed filter bank coefficients used in factorisation. Nevertheless, signal level synthesis was used in the included publications due to employment of several back-end feature extractors not directly compatible with the internal NMF feature space, and to produce wave files for computational and subjective quality evaluations.

In robust ASR, the benefits of properly functioning speech enhancement are obvious. Any noise features in the input will make it less speech-like, thus a worse match to back-end models representing speech. Not even multi-condition training can truly compensate the mismatch, especially if the noise is non-stationary and unpredictable in its behaviour. Whenever the gains from noise removal outweigh

Figure 5.1: Steps taken by a factorisation and enhancement system to compute spectrogram estimates, a time-varying filter, enhanced speech features, and sparse classification output.



the possible loss of actual speech features, improvements in back-end recognition accuracy can be expected. The proposed NMF algorithms have yielded uniform increments in recognition rates even using semi-supervised enhancement without a trained noise model [P4, P6]. As usual, re-training the back-end with similarly enhanced speech will reduce the mismatch further [P8]. Apart from ASR, speech enhancement has applications in recording, transmission and storage for better intelligibility, more efficient compression via reduction of noise-like information, and in any speech processing that benefits from a cleaner input signal.

5.3 Sparse Classification

Sparse classification (SC), briefly introduced in Section 2.4.3, is an alternative approach to ASR exploiting the factorisation output without converting it back into a spectral or waveform domain. The nomenclature arises from *sparse representations*, where enforcing sparsity on the model has been found beneficial for discovery of key features from noisy or mixed data [26]. Although sparsity constraints regularly appear in separation and enhancement tasks, in SC they can be considered almost essential [48] unlike in separation, where the bias introduced by sparsity objectives may be even detrimental for quality metrics [189].

As seen in the diagram of Figure 5.1, sparse classification output is derived from activation weights of factorisation with no need to construct the spectrogram estimates. In simple tasks with few discrete output classes such as keyword spotting or speaker recognition it suffices to observe, which atoms were activated in factorisation. Assuming that activations can be represented as a fixed-length vector, common classification algorithms are applicable for training and evaluating the class borders [143]. For longer observations, where the temporal structure of words or other events is important, it becomes necessary to model the temporal dimension as well.

Early sparse classifiers used histogram modelling for multi-digit recognition tasks, assigning state likelihoods uniformly over the duration of atoms and windows [42, 178]. Notable temporal blurring was consequently present in the likelihood estimates over utterances. It was soon found out that assigning state transcription on a frame level to multi-frame atoms improved the decoding accuracy significantly [190]. For more general speech recognition tasks with a large vocabulary, explicit state modelling over time becomes effectively essential.

In this thesis' work, sparse classification is based on assigning *label matrices* to speech atoms. Let us assume that the language model employed in recognition contains Q states, which may denote e.g. phonetic or sub-word models. Each speech atom, whose spectral feature content is a $B \times T$ matrix, is also given a $Q \times T$ label matrix, reflecting the state membership of each frame. The matrix may be binary, i.e. only one state entry per frame is active at weight 1, or fuzzy so that several state candidates may be active at variable weights. After determining the speech activations, the same reconstruction formulae that are used for spectrogram estimation in (2.2) and (4.2) for producing a $B \times T_{\text{utt}}$ spectral feature estimate, are applied to label matrices yielding a $Q \times T_{\text{utt}}$ matrix of state weights over observation frames. The matrix thus conveys similar information as likelihood estimates from conventional evaluation of GMMs for the frames of an utterance.

Although the distribution and magnitude scaling of SC state weight estimates will be different from the output of GMM evaluation, they can be decoded using HMMs trained with a conventional back-end as long as the correct path is found by the Viterbi algorithm. For AURORA-2 data, 96–97% digit recognition accuracy was achieved for clean speech already with early versions of the SC framework [P1, 44]. Keyword accuracy for 1st CHiME clean development data was approximately 93% [P3]. In both cases the error rate is slightly higher than for baseline GMM evaluation. A likely reason is that the presented SC systems operate on plain mel spectra, which are not as accurate as mel-cepstral features for classifying phonetically close keywords like 'five'/'nine' or 'b'/'v'.

In noisy conditions, the SC approach has repeatedly surpassed conventional GMM recognisers due to its superior robustness via explicit noise modelling [P6, 44]. However, compared to a GMM back-end with NMF feature enhancement and model re-training, there have been results favouring either FE or SC, depending on the NMF model and back-end parameters [P6, 44]. Another significant factor is the method of assigning the label matrices. The first SC systems used binary matrices acquired simply by assigning the single state determined by forced alignment as the only active entry of atom-frames. Thereafter more advanced algorithms such as ordinary and partial least squares regression (OLS, PLS) and NMD learning have been proposed with solid improvements on recognition accuracy just by better translation of activations into state estimates [70, 106].

Although the direct SC approach described here is unlikely to provide a complete replacement for GMM evaluation, obviously the information available in speech activation weights is meaningful, hence it should not be ignored in decoding. Similarly to other exemplar and template systems [17, 21, 145, 146, 164], the information should be exploited even more efficiently with integration to other recognition methods. For example, FE and SC streams have been found complementary in multi-stream recognition [204]. Other combinations of methods, both previously proposed and emerging, are discussed in the next section.

5.4 Combined Recognition Methods

In ASR, there are standardised processing chains like GMM-HMM recognition with cepstral features, which are readily available in software implementations [213]. Despite their shortcomings in some applications like robust recognition, they have been fine-tuned over years to cover several algorithmic stages like language modelling, feature extraction, statistical modelling and so on. Reinventing the whole toolset would be a daunting task and usually not even necessary. Therefore new paradigms are often tested with replacement or combination of new and established system components. This also applies to NMF-based processing, which may act in multiple roles ranging from plain front-end enhancement to direct state likelihood estimation or word spotting. This section illustrates a few approaches proposed for joint recognition with NMF as one of the components involved in more complex systems.

The terminology of alternative modelling methods can be derived from parallels in artificial neural network (ANN) systems, especially multi-layer perceptrons (MLPs), which started to emerge for ASR in the 80s [100], and then have repeatedly appeared in literature since the 90s [19, 116, 141], gradually evolving into deep neural networks (DNNs) [67]. There are two major branches of systems employing MLPs in ASR. In a *hybrid* approach, an ANN is trained to produce direct posterior likelihoods for HMM states, thus replacing the whole statistical modelling and evaluation [5]. In *tandem* systems, the ANN outputs are modelled statistically with GMMs, hence acting as features instead of e.g. mel-cepstra [64] but not producing direct state likelihoods. Further variants, comparisons and insights to these major routes are provided in later work [19, 176].

The plain sparse classification system described in this thesis is essentially hybrid-like, because it produces state likelihoods as its output. However, in [170], the SC output is modelled with GMMs, making the system similar to tandem recognition. These parallels in ANN- and NMF-based single-stream recognition are illustrated with simplified flowcharts in Figure 5.2. The first two paths a) and b) represent conventional GMM evaluation of e.g. MFCC or PLP features, optionally with an enhancement front-end. The middle paths c) and d) correspond to hybrid and tandem recognition with ANNs, respectively. The last two paths represent direct sparse classification and statistical modelling of SC outputs.

Single-stream processing in consecutive algorithm steps is not the only option for recognition, though. In [167, 168], a dynamic Bayesian network (DBN) is used to combine SC and MFCC likelihoods. Similarly in [169], estimates from SC and a three layer MLP are combined either by summing or multiplying the state probabilities to produce the combined posterior probability for decoding. In [40], SC and NMF-enhanced MFCC probabilities are combined with a product rule. In [203, 208], a *bi-directional long short-term memory recurrent neu-*

Figure 5.2: Main components of single-stream recognition paths employing statistical modelling, spectrogram factorisation, and neural networks, starting from spectral features and ultimately producing likelihoods for back-end decoding:
a) conventional GMM system with e.g. mel-cepstral features [133]
b) GMM system with NMF feature enhancement [44, 137]
c) hybrid ANN recognition [5]
d) tandem ANN recognition [64]
e) NMF sparse classification [37, 44]
f) statistical modelling of SC output [170]

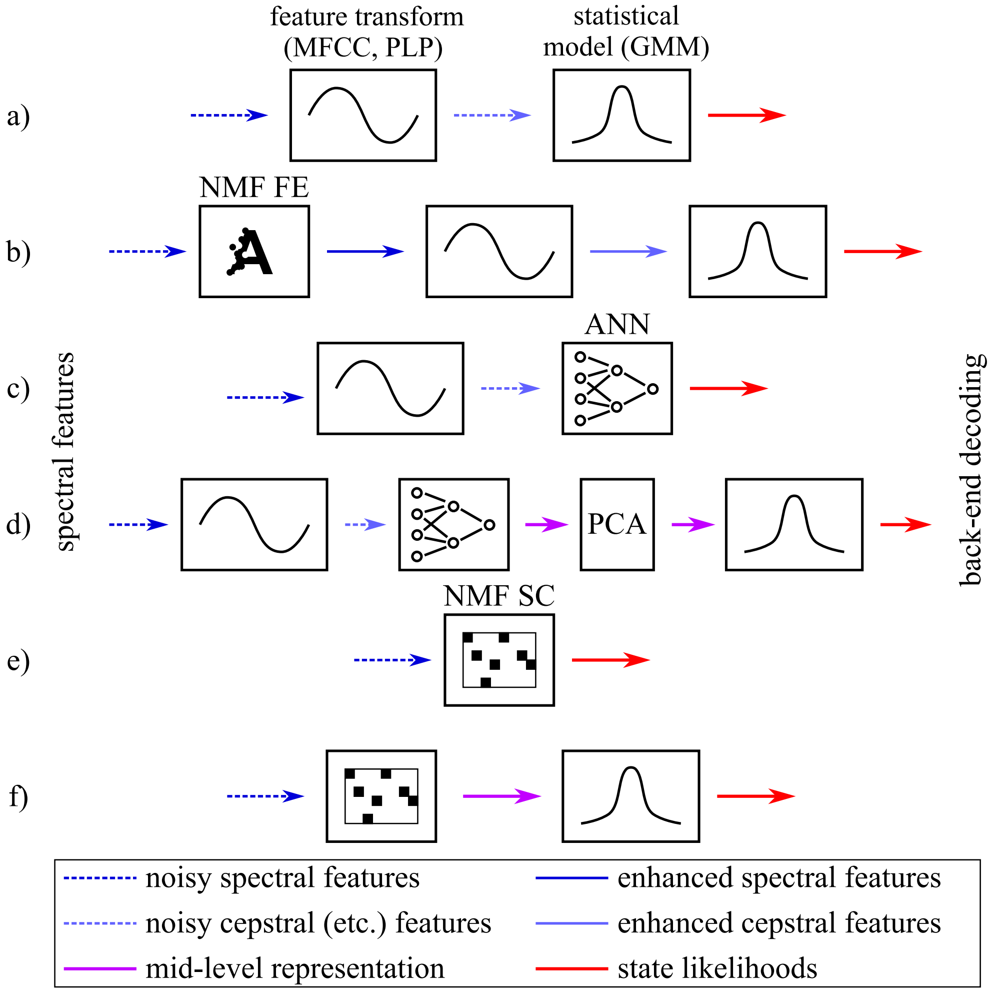
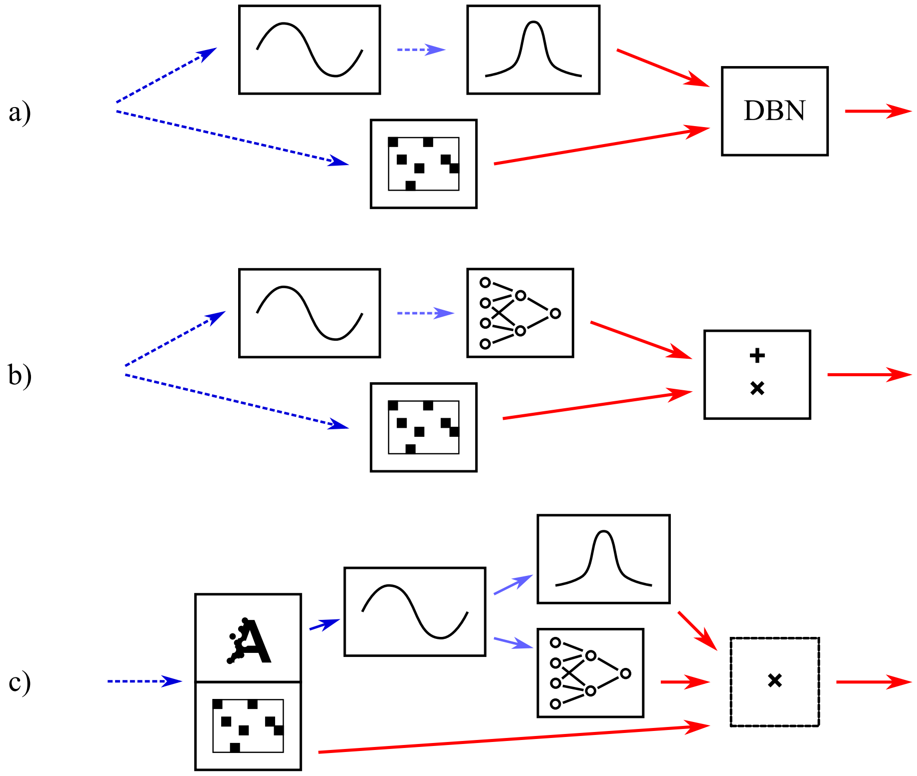


Figure 5.3: Recognition paths employing stream combinations:
a) MFCC+SC with DBN combination [167]
b) ANN+SC with sum or product likelihood combination [169]
c) triple-stream MFCC+ANN+SC recognition from enhanced features [35, 204]



ral network (BLSTM-RNN) is used in conjugation with NMF-enhanced MFCC-GMMs for probability combination. In [35] and [204], three streams are combined; NMF-enhanced MFCCs, sparse classification, and a BLSTM-RNN. These latter systems, typically computing the product of stream probabilities with exponent weight factors, can be referred to as *multi-stream*, hybrid-like recognisers. Figure 5.3 shows schematic views of three of these systems, namely [167], [169] and [204]. Other combinations can be illustrated similarly or as subsets of these examples.

In these multi-stream experiments, all feature streams have been found complementary, that is, combined evaluation surpasses the recognition rates of its single components even if the FE and SC outputs are derived from the same NMF system. Apart from FE and SC streams, NMF output has also been used for estimating masks in uncertainty and missing data decoding [44, 78, 79]. Meanwhile, deep neural networks have gained a lot of attention in ASR, being employed in

ASR applications by major companies and producing state-of-the-art results in recognition of real-world speech [67]. They should be able to provide even more complementary information to joint systems, again improving the overall recognition rate. Yet another path beyond the scope of this thesis are spatial algorithms, which are likely to become increasingly important as multi-microphone devices gain popularity, and demonstrably improve the recognition results further [16, 125, 175].

For actual combination of streams, many more algorithms have been proposed in literature than the DBN, sum, and product approaches previously employed in joint systems containing NMF components. *Recogniser output voting error reduction* (ROVER) uses a variety of voting schemes to find an optimal word transition network from multiple system outputs [32]. *Confusion network combination* (CNC) is its later extension [29]. BAYCOM stands for *Bayesian combination* using a decision-theoretic approach that is expected to provide optimal combination weights even for streams with considerably differing error rates [147]. *Driven decoding algorithm* (DDA) performs dynamic search between a primary system and auxiliary systems or manual transcripts [88]. There is no particular reason preventing the use of NMF-based components in these fusion methods as well, although no examples appear to exist in literature yet.

We can conclude that there is a multitude of established and novel recognition paths, NMF-based or not, which provide partially overlapping yet ultimately complementary information for joint recognition. This raises interesting questions on how to incorporate the strengths of different methods in a joint system while minimising the redundancy and computational complexity. Because the best performing robust systems are currently relatively heavy combinations of multiple methods [3, 183], these questions can be expected to remain highly relevant in the quest for human-like or even superhuman ASR performance.

5.5 Other Recognition Tasks

5.5.1 Speaker Recognition

As discussed in Section 4.3, long-context exemplars and templates are able to model discriminative spectro-temporal characteristics of individual speakers. Combined with the robustness of additive multi-source models, sparse classification via NMF can be used for *speaker recognition* and *verification*, the former standing for determining the correct profile among a set of speakers, and the latter for confirming whether the observed voice matches a certain profile.

There have been earlier examples of sparse methods for speaker recognition. Sparse representations without non-negativity constraints nor a noise model were proposed for speaker recognition in [86], [120] and [177]. NMF coefficients derived from GMM mean supervectors projected into a non-negative space were used in [104], and supervised NMF with a pre-trained speech basis in [74]. Non-negative tensor factorisation of cortical features was proposed in [209]. However, none of these systems employed explicit noise modelling in factorisation of spectrograms. Instead, general robustness was achieved with feature selection and sparse coefficient extraction.

The NMF approach with long temporal context and a dedicated noise basis was first demonstrated in [143]. Template bases were constructed for the 34 GRID/CHiME [3] speakers as in [P6]. Activation weights from factorisation with combined speech and noise bases were then used to estimate the active speaker. Multiple classification algorithms were used, starting from simply observing the maximum total activation weight of each speaker's basis, and then advancing into inner product scoring, probabilistic linear discriminant analysis (PLDA), and sparse discriminant analysis (SDA). The latter variants produced under 0.5% error rate in clean or near-clean conditions, and 5.0% average error rate over noisy conditions from +9 to -6 dB, uniformly surpassing more conventional HMM and GMM-UBM (universal background model) algorithms. In [P5], a group sparsity constraint was introduced to improve the discriminative capability of the system, reducing the average noisy error rate down to 4.3%. In [P4], speaker identity estimates are used for choosing the best front- and back-end models for ASR of unknown speakers via feature enhancement. In [P8], newly introduced speakers are approximated with existing training profiles.

The results show that NMF with a long temporal context is highly viable for speaker recognition and approximation even in very noisy conditions, unlike short-context algorithms commonly employed for clean speech. One open issue with the NMF approach is that many speaker recognition applications also require robustness against channel variations, which is not achieved with the proposed system. However, preliminary results suggest that channel invariance can be implemented in the NMF framework too, further increasing its applicability [43].

5.5.2 Voice Activity Detection

No natural speech is completely continuous. Unless utterances have been tightly cropped in preprocessing, there will be segments without voice activity. In noisy conditions this does not equal silence, though, because other sources may be active at amplitudes comparable to target speech. *Voice activity detection* (VAD) or *speech activity detection* (SAD) stands for locating speech segments from a noisy input. Its applications include selecting speech segments for further processing and storage, determining silence state likelihoods for back-end models, updating noise models, and activating speech- or noise-specific system components.

Due to potentially high energy levels of noise events, just observing the input signal activity does not suffice for robust VAD. More advanced algorithms employ e.g. subband analysis, modulation features, spectral subtraction, phoneme recognition, and statistical models [108, 144, 215]. In [20], statistical modelling of a sparse (K-SVD) representation of speech was used for VAD, although without a separate noise dictionary. However, similarly to speech recognition, in noisy conditions it becomes increasingly beneficial to model noise explicitly. The proposed NMF framework does this, and reveals approximate energies of speech and noise via activation weights of their corresponding bases. In [P1], [44] and [212], SNR estimates and silence state weights were derived by comparing speech and noise activations. In [P4], a voice activity estimate was acquired by convolving speech activations with grammar-dependent weight profiles. In each case, robust VAD was achieved even in environments containing non-target voices and babble noise. Obviously NMF-enhanced signals or features can also be fed to an established VAD algorithm for another detection path.

Especially with an explicit noise model, sparse representations appear viable for voice activity detection, being able to model concurrent speech and noise events individually. The reliability of detection generally depends on the same factors that affect separation quality with model accuracy or possible mismatch playing a major role. Apart from better speech and noise modelling, NMF-based VAD would benefit from incorporating better classification schemes to the interpretation of activations, similarly to SC-based speech recognition, where more accurate label assignment has improved the accuracy significantly without even changing the factorisation output [70].

5.5.3 Further Applications

Apart from the described common speech processing tasks, sparse representations generalise to other detection and classification problems as long as appropriate source models can be acquired. One frequently appearing real-world problem is detecting overlap in multi-speaker scenarios like meetings. The additive model of NMF is inherently fitting for this purpose as well, and has been proposed in literature [34, 186]. In [2], speaker age and gender are estimated from a GMM-based supervector. In [154], non-linguistic vocalisations are detected among speech. In [15, 49, 61], sparse representations are used for sound event classification. Even more applications arise when the scope is extended to e.g. music. In short, sparse classification with NMF has potential for several speech and audio processing tasks, especially in scenarios involving multiple concurrent target speakers or noise sources. The model also facilitates extraction of several types of information simultaneously as demonstrated in [P4], where voice activity, speaker identity, and speech content were all observed within the same framework.

Chapter 6

Current Performance and Practical Considerations

The whole speech processing framework described in this thesis aims at solving a practical problem, namely extracting the lexical content or other information from speech signals corrupted by competing sources and other real-world phenomena. Although maximal recognition accuracy is obviously a definite goal, eventual implementations must also be applicable to everyday tasks and situations using common devices. This chapter provides a brief overview to results achieved with proposed factorisation-based methods, and discusses their practical aspects including modelling, storage, and computational issues.

6.1 Quality Measurements

The integral question concerning any ASR system intended for real-world use is "Does it actually work?" The ultimate answer could only be determined by deploying the system to a widely used application and measuring the success rate in its actual purpose. Because the described NMF framework has not reached this stage yet, and more generally there is little information available concerning the use of NMF-based systems in practical applications, current results are derived from evaluations using public test databases. While these tasks are typically simplified and limited in their scope, they provide a standardised benchmark for assessing the performance of different approaches and implementations. A summary of the main characteristics of employed databases and a list of included publications where they appear is given in Table 6.1. Principal results for these databases are discussed in this section, while more detailed comparisons between algorithm variants and competing methods can be found in the publications.

Table 6.1: Main characteristics of noisy speech databases used in experiments.

database	vocab. size (type)	noise	SNRs	used in
AURORA-2	11 (digits)	8 types	$+20 \dots -5$ dB	[P1]
CHiME/GRID	51 (commands)	living room	$+9 \dots -6$ dB	[P2]–[P7]
CHiME/WSJ	5000 (newspaper)			[P8]

6.1.1 AURORA-2 Speech Recognition Rates

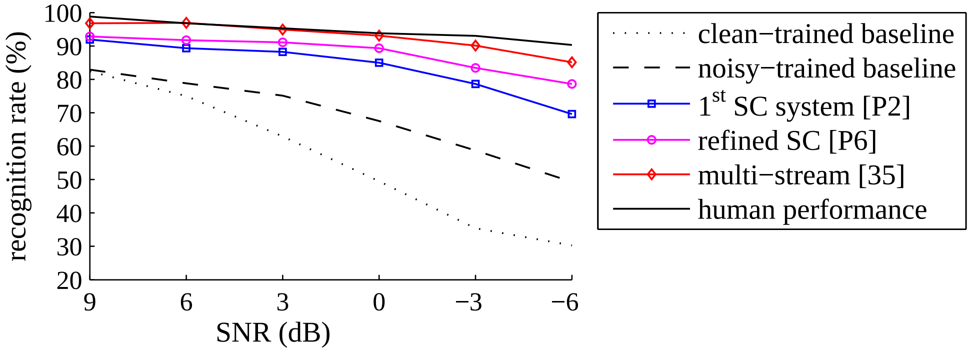
The first systems covered by this thesis were evaluated on AURORA-2 data, derived from the TIDigits connected digit recognition task [95]. In this corpus, sequences of 1–7 spoken digits are mixed with multiple real-world noise types at SNRs from +20 to -5 dB [68]. The framework described in [44] and [190] employed three recognition paths; missing data masks, feature enhancement, and sparse classification, which were compared to reference results from imputation and a multi-condition trained baseline recogniser with mean and variance normalisation. This initial version performed reasonably well, but only surpassed the multi-condition trained GMM recogniser in conditions with heavy, matching noise [44]. One major reason was the system's worse initial classification accuracy in clean conditions. Another was limited re-training in the paths using an external recogniser. The accuracy of sparse classification improved in [P1], where convolutive modelling was introduced. Further optimisations [40, 48] and system combinations [40, 169, 170] managed to address the bottlenecks to the point that the latest published results are among the highest reported for AURORA-2 [40]. Current word error rates over 20–0 dB are 3.1% and 4.7% for test sets 'A' and 'B', respectively, using a multi-stream SC+FE system.

6.1.2 CHiME Speech Recognition Rates

A majority of the work covered by this thesis has been conducted on the 1st CHiME Challenge data [3], or the closely related Track 1 of the 2nd CHiME Challenge [183]. These corpora are derived from GRID speech [14], whose small vocabulary command utterances are mixed with household noise at SNRs from +9 to -6 dB. Although the speech recognition task is again heavily simplified, the noise environment in CHiME data is notably varied and highly non-stationary.

In the 1st CHiME Challenge, the submitted system [P2] performing sparse classification with large exemplar bases ranked approximately 6th of the 13 participants. The enhancement-based variant was not as efficient, primarily because it used the clean-trained baseline back-end [45]. Later refinements, especially introduction of optimised spectral features and temporal deltas [P3], and a multi-

Figure 6.1: Development of CHiME/GRID keyword recognition accuracy compared to baselines. Black lines correspond to clean- and noisy-trained GMM baselines and human performance. Coloured lines show results for the 1st CHiME workshop SC entry [P2], a refined SC system with temporal dynamics [P6], and finally the multi-stream entry to the 2nd CHiME workshop [35].



condition trained back-end boosted the performance of both SC and FE significantly [P6]. Finally, the SC and FE streams were combined with a neural network component in [204]. This triple-stream system ultimately produced the best results for Track 1 of the 2nd CHiME Challenge [35]. The development is illustrated in Figure 6.1, where the systems from [P2], [P6] and [35] are compared to GMM baselines and human performance. Note that the comparison includes results from the 1st and 2nd GRID-based CHiME tasks, which are almost but not exactly identical. Nevertheless, the overall development should be apparent. For this small vocabulary task, the multi-stream system is already approaching the robustness of human ear.

As a general trend in robust ASR, the highest-ranking systems for any corpus tend to be combinations of multiple components and streams. This was also the case for the WSJ-based Track 2 of the 2nd CHiME Challenge, where the proposed NMF system is notably effective as an enhancer [P8], but does not match the overall performance of systems employing heavily customised back-ends [183]. However, a combined approach with a neural network already provided major improvements [35], and an optimised version of the joint framework produced the highest reported results for the corpus in 2014 [36].

6.1.3 Enhancement and Separation Quality

Whereas the previously described benchmarks were based on keyword scoring in ASR tasks, there are also alternative measures for separation and related problems. Plain separation quality can be compared with objective metrics like signal-to-distortion ratio (SDR) or estimated perceived quality [24, 28, 58, 117, 132, 182], or with subjective evaluation [58, 71, 117]. In [87], three 1st CHiME systems were compared using computational metrics and listening tests. The proposed NMF system [P6] produced the best results regarding SDR and other objective metrics, although not uniformly. It also passed a t-test of being significantly ($p = 0.05$) better than the alternative methods and unprocessed signals in a listening test. The average SDRs measured with BSSeval toolkit [182] over complete noisy sets have been approximately 7.0–8.8 dB for compact modelling schemes and 9.5 dB for large exemplar modelling, compared to -0.7% for the unenhanced signals [P7].

6.1.4 Speaker Recognition

In speaker recognition, the NMF framework has not been implemented for the most popular evaluation corpora yet. Furthermore, many evaluations give more emphasis to robustness against channel errors and mismatch than additive noise. Nevertheless, in [143] and [P5], the preliminary NMF system clearly surpassed established speaker recognition algorithms in noisy conditions using the 1st CHiME data and its 34 speakers for evaluation, achieving 99.7% accuracy at 9 dB SNR and 95.0–95.7% average accuracy over +9 . . . -6 dB. Because early results appear promising, it would be desirable to implement the system for other corpora to learn more about its capability in speaker recognition and verification tasks.

6.2 Modelling and Efficiency Issues

6.2.1 Model Adaptivity

Early examples of speech processing via NMF demonstrated separation of two known speakers with pre-trained speech bases [153, 158], or denoising by including a trained noise basis [158]. Similarly the exemplar-based framework first employed sampled speech and noise bases in supervised separation and classification [44]. However, in all these experiments two major assumptions were made in training, namely that

1. sufficient training material is available for all sources, and
2. speech and noise profiles remain approximately fixed.

Neither of these assumptions can be expected to hold universally true in general purpose robust processing. Therefore one recurring theme in this work is improving the system's adaptivity to changing conditions.

Concerning noise models, the nearby context of utterances has been exploited to an increasing degree [P2, P6, P8, 41, 185]. In addition, [P4] removed the assumptions of annotated utterance locations and bi-directional context. Instead, voice activity was estimated with NMF, and the noise model was updated continuously from non-speech segments. Even in a critical scenario, where neither training data nor context is available for learning a noise model, *semi-supervised* separation has been demonstrated [P6, P8, 7, 199]. In this variant, only a speech basis is trained beforehand, whereas the noise basis is updated from the noisy observation. Although there is a considerable risk of overadaptation and thus losing speech features, with cautious parameter selection uniform improvements have been observed both in enhancement and in direct classification [P6].

Regarding speech models, as stated in Sections 3.2 and 4.3, the spectro-temporal behaviour of speech is more constrained than the enormous variety of noise sources. An averaged or mixed speech model can be constructed for speaker-independent factorisation, if training a matching speaker-dependent model is not viable. Nevertheless, a closer approximation always benefits separation, especially in scenarios involving overlapping non-target speakers. One promising direction for adaptive speaker modelling is constructing a multi-speaker basis with group sparsity constraints, favouring an output where a small set of closest matching models approximates the target speaker [P5, P8, 165]. Alternatively, in [7] a generic speech basis is permitted to adapt into a new speaker's profile with constraints on its spectral divergence from the original content. With properly defined cost functions, the method appears viable for adaptive enhancement of speech from an unknown speaker, albeit it may be less suited for sparse classification as the spectral content of speech atoms may change during factorisation.

Another partially open issue in speech modelling is representing a large vocabulary with robust long-context units, while keeping the model size manageable. Potential approaches include variable-length units and clustering algorithms [P7, P8], and HMM-driven factorisation discussed in Section 4.1.5. From a wider perspective, combining NMF with neural network components or their concepts appears highly promising [35, 169, 170, 204].

The last major factor in overall adaptivity is robustness against channel errors and mismatched response. For limited bandwidth or other loss of features, missing data techniques and imputation provide a potential solution [38, 39, 44, 80]. If the channel's response is distorted instead, it is possible to estimate a compensation filter within the factorisation framework [43]. Finally, an auditory-motivated feature representation should be considered for achieving human-like general robustness against unpredictable deviations in the signal [101, 109, 162].

6.2.2 Data Requirements and Model Sizes

Results from separation and recognition experiments have repeatedly confirmed that the factorisation-based framework benefits from long temporal context and overcomplete modelling with a comprehensive dictionary of diverse speech patterns, sound events, and their variants [P6, 44]. However, an inherent drawback of explicit segment models is their rapidly increasing size in comparison to statistical models. Direct consequences include greater memory consumption and computational complexity, but also higher requirements for source data in order to acquire a comprehensive basis of specialised long-context patterns.

Early small vocabulary classifiers for AURORA-2 [P1, 44] and 1st CHiME [P2, P6] corpora used large exemplar bases with 4000–5000 atoms for both speech and noise. Further experiments revealed that even larger bases still improve the results in the proposed exemplar framework [40, 48]. The practicality of such models could be disputed, though. For example, the best performing system in [40] employs 14 000 atoms, each a 30-frame segment in a 23-band spectral space. Together they contain 9.66 million entries and require almost 40 megabytes as single precision float variables. In [P3] and [P6], even more spectral bands with delta and stereo features were employed, taking the overall memory allocation to hundreds of megabytes. Because factorisation involves iterative matrix operations on the complete arrays, processing a short utterance could take several minutes of CPU time. These figures appear high for a small vocabulary task exploiting speaker-dependent models. On the other hand, randomly sampled exemplar bases contain a lot of redundancy, thus significantly smaller model sizes can be achieved via better acquisition algorithms. In later work, multiple directions were studied for making the models more efficient by finding a compact representation of relevant patterns while trimming the redundant and unused parts of models.

In speech modelling, already the system presented for 1st CHiME data in [P2] contained some optimisation over completely random selection, namely reducing the overrepresentation of the most frequently appearing words. However, more drastic reductions were proposed in [P6], where exemplar models were replaced with speech templates, each modelling a sub-word pattern as an average of training instances. The size of speech and noise bases was reduced from 5000 to 250 atoms, standing for a 20-fold reduction in data size and computational complexity. For current hardware and NMF algorithms, this kind of difference may define whether the factorisation can be performed in real time. Despite the heavy compression of source models, only modest reduction was observed in enhancement quality. The losses in sparse classification were greater, though, mainly because the slightly blurred mel-spectral templates are not as reliable as classifiers as exact speech exemplars. In [P7] and [P8], the concept was taken even further by introducing variable-length templates, where characteristic patterns were detected and modelled in a descending order of length. Improvements were observed compared to fixed-length templates for the 1st CHiME data [P7]. Finally, even larger savings could be achieved by using HMM-driven frame [53, 115, 119] or segment [46] models, which provide a flexible way to represent acoustic trajectories and temporal deviations, thus circumventing the rigidity of exemplars and templates.

In noise modelling, the diversity of noise events means that a generic basis can never be simultaneously small and accurate. Therefore the key to efficient noise models is high adaptivity so that only patterns relevant for the current environment are retained in the basis at any given time. In [P2] and [P6], noise atoms were selected exclusively from the context of noisy utterances. The experiments in [P6] revealed that even a small noise basis can still provide decent separation quality, especially with the convolutive model. In [P4], a continuously updated noise basis with aggressive pruning was introduced for processing long inputs adaptively. The smallest model sizes and training data requirements are achieved with *semi-supervised factorisation*, where a small basis is updated during factorisation of utterances [P6, P8, 199]. However, thus far its separation quality has been inferior to variants employing a separate noise basis due to the risk of capturing concurrent target speech patterns [P6, P8]. Thereby maintaining a dedicated noise model is still recommendable. For improving its performance, NMD learning [P8, 208], and possibly variable-length methods similar to proposed speech clustering appear viable. Despite the diversity of noise events, there is also a lot of local redundancy in noise due to stationary sources and recurring events, hence the characteristic events can often be compressed into a compact model.

6.2.3 Computational Complexity

A persistent challenge in practical use of NMF algorithms is the lack of closed-form solutions, meaning that iterative descent algorithms are used for finding an approximate solution. The multiplicative update rules introduced by Lee and Seung [91, 92] are able to find a passable approximation for e.g. enhancement filters in only a few [77, 202] or maybe tens of iterations [75, 158], although concerning the cost function the solution still improves gradually. Hundreds of iterations are commonly used, especially in supervised factorisation [44, 158, 199]. When a sparsity constraint is introduced, the spectral estimate usually converges faster than the sparsity of activations, likely because a steep initial descent can be found in the spectral cost, while the contribution of sparsity to the total cost is considerably smaller. In [48], increasing the iteration count from 200 to 300 was found beneficial for sparse classification, and in [40] up to 600 iterations still improved the accuracy in both FE and SC. However, the computational complexity of NMF and NMD increases linearly to iteration count, basis size, feature space dimensionality, and window length. Especially for large exemplar bases the costs quickly become nontrivial and unfeasible for real-time systems. Consequently, several paths have been studied for addressing the computational costs of NMF.

The first logical route is reduction of model sizes, which depend on the system's spectro-temporal resolution and atom count. Reducing the spectral resolution is possible, although already the current mel-spectral representation is quite compact, and from the separation point of view there is actually an incentive to increase it. Lower frame rate is viable for enhancement, but not so much for sparse classification. Total window length is heavily tied to separation quality, thus reducing it means a major trade-off unless alternative structures like HMMs manage to replace the window model. These parameters were evaluated in [75] with a goal of achieving real-time performance in NMF-based enhancement.

Apart from data dimensions, the other major factor in computational complexity are the solving algorithms themselves. Recently there have been examples of replacing the established gradient descent rules with their quadratic counterparts [179, 191, 216]. In [191], an *active-set* approach is also employed, minimising the number of activation coefficients being updated. For NMD, an online algorithm based on sufficient statistics has been proposed for piecewise processing of otherwise inconveniently large basis updates [196, 197]. In [161], low-rank NMF is used for real-time speech denoising.

Finally, NMF solving can be speeded up simply by improving the efficiency of low-level operations. There are toolkits like OpenBlissART [200, 201] and FASST [125] implemented in C/C++ with internal optimisation of the solving algorithms. Another significant trend is parallel computing with multi-core CPUs or many-core general purpose graphics processing units (GPU, GPGPU). The large matrix multiplications appearing in NMF algorithms are well suited for parallel computing. A GPU implementation was used in [48], and nowadays it can be found in OpenBlissART as well [201]. Parallelisation of NMD has thus far appeared more difficult due to its three-dimensional basis structure and stepwise looping over at least one dimension in standard implementations, but optimised versions are gradually appearing for convolutive updates too.

6.3 Summary

In short, spectrogram factorisation with long-context atoms appears potent in separating speech from difficult noisy mixtures, and in classification of phonetic or speaker-related information under noisy conditions. Still, for maximal performance it is beneficial to incorporate further information and system components from alternative separation and recognition methods [40, 46, 169, 204]. From a practical point of view, there is obviously a trade-off between quality and computational costs, hence the system should be optimised regarding both objectives for better efficiency considering resource allocation. For this kind of optimisations, the model reduction schemes proposed in [P6, P7, P8] and adaptive continuous input processing presented in [P4] should be studied in more depth.

Chapter 7

Conclusions and Future Work

In this thesis, non-negative spectrogram factorisation algorithms were applied to robust processing of speech. The fundamental concept in the presented work is modelling a mixture signal as an additive combination of spectrogram components belonging to constituent sources, thus separating the mixture into single-source feature streams and further to their characteristic events. As a separating front-end, the methods are commonly used for enhancement of noisy speech for improved perceptual quality, transmission, and automatic speech recognition (ASR). In addition, the system acts as a classifier, revealing information on the contained phonetic patterns, sound events, and speaker identities.

The foremost strength of the approach lies is its ability to model multiple concurrent sound sources explicitly in complex multi-source audio scenes. Speech features can thus be recovered in difficult conditions, where they are heavily masked by varied noise events. Similarly the methods are applicable e.g. for separation and classification of multiple overlapping speakers. Recurring concepts in the employed factorisation algorithms include long temporal context for modelling the spectro-temporal behaviour of sound events, and sparsity for finding a small set of best matching components to model the mixture. Main scientific contributions of this work include refinement of spectral feature spaces, new basis acquisition algorithms, combination of multiple speaker-dependent models, improved adaptivity to varying noise conditions, and moving toward robust large vocabulary speech recognition with factorisation-based methods.

The described speech processing framework built around factorisation algorithms performs several tasks, including feature extraction, speech and noise model acquisition, actual separation of underlying sound sources, and finally their enhancement or classification for extracting the relevant information for the task. Each stage of the overall process is addressed in the contained publications. Apart from improving separation and recognition quality, the conducted work aims at finding more efficient models and algorithms for practically viable implementation, and higher adaptivity to the great variety of scenarios and changing conditions in real-world environments.

The main focus of the work has been on improving the performance of speech recognition in realistic conditions, where conventional algorithms have struggled to the extent that widespread adoption of ASR systems has repeatedly been delayed. The described algorithms and modelling methods have shown strong performance especially in scenarios involving heavy, non-stationary noise, which can be represented explicitly using the proposed models. In addition to standalone recognition, the methods have been combined with alternative recognition paths, providing complementary information and eventually achieving state-of-the-art results in public ASR evaluations. The framework has also been successfully used for robust speaker recognition, voice activity detection, and sound event classification.

Regarding future work, there is still room for improvement in each of the previously described stages of the overall framework. In basis acquisition, higher flexibility, adaptivity and automation would be desirable for extracting efficient speech and noise models from diverse inputs. Alternative feature representations such as auditory-inspired or phase-sensitive multi-channel features are potentially able to improve the separation quality and robustness of the approach. In factorisation, there is a large variety of structures and priors available for incorporating information that is not exploited by the current methods. The computational implementation of algorithms is also under ongoing research for reducing the costs of factorisation. Finally, deriving the enhancement, recognition, and classification results from the factorisation output is a delicate task, where multiple paths and their combinations are currently used. This raises the question whether currently used parallel streams can be merged into a unified system combining the strengths of each component. Especially the renewed interest in neural networks and their combination with sparse representations in ASR is likely to remain a central topic in future research.

To recap the current state of robust speech processing, the complex behaviour of human speech and even greater diversity of environmental sounds form a difficult problem, where human hearing still outperforms computational methods. Nevertheless, gradual algorithmic improvements, huge resources, and sharing of ideas within the research community have reduced the gap to the extent that ASR applicable to everyday tasks is firmly moving from the realms of fiction to reality. While sparse representations and spectrogram factorisation alone are unlikely to solve the complete range of robust speech processing problems, they have repeatedly produced strong results, and a lot of their potential is undoubtedly still to be revealed. The presented work has hopefully offered valuable insights to the topic, and motivated further work to eventually reach the long-awaited goal of fluently listening and understanding machines.

Bibliography

[1] G. Aimetti. “Modelling Early Language Acquisition Skills: Towards a General Statistical Learning Mechanism”. In: *Proceedings of the Student Research Workshop at the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL SRWS)*. Athens, Greece, 2009, pp. 1–9.

[2] M. H. Bahari and H. Van hamme. “Speaker Age Estimation and Gender Detection Based on Supervised Non-Negative Matrix Factorization”. In: *Proceedings of IEEE Workshop on Biometric Measurements and Systems for Security and Medical Applications (BIOMS)*. Milan, Italy, 2011.

[3] J. Barker, E. Vincent, N. Ma, C. Christensen, and P. Green. “The PASCAL CHiME Speech Separation and Recognition Challenge”. In: *Computer Speech & Language* 27.3 (2013), pp. 621–633.

[4] L. Benaroya, L. M. Donagh, F. Bimbot, and R. Gribonval. “Non Negative Sparse Representation for Wiener Based Source Separation with a Single Sensor”. In: *Proceedings of the 28th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Hong Kong, 2003, pp. VI 613–616.

[5] H. Bourlard and N. Morgan. *Connectionist Speech Recognition: A Hybrid Approach*. Boston, MA, USA: Kluwer Academic Publishers, 1994.

[6] J. J. Carabias-Orti, F. J. Rodriguez-Serrano, P. Vera-Candeas, F. J. Cañadas-Quesada, and N. Ruiz-Reyes. “Constrained non-negative sparse coding using learnt instrument templates for realtime music transcription”. In: *Engineering Applications of Artificial Intelligence* 26.7 (2013), pp. 1671–1680.

[7] M. A. Carlin, N. Malyska, and T. F. Quatien. “Speech Enhancement Using Sparse Convolutive Non-negative Matrix Factorization with Basis Adaptation”. In: *Proceedings of the 13th Annual Conference of International Speech Communication Association (INTERSPEECH)*. Portland, OR, USA, 2012, pp. 582–585.

[8] B. Chen, S. Chang, and S. Sivadas. "Learning Discriminative Temporal Patterns in Speech: Development of Novel TRAPS-Like Classifiers". In: *Proceedings of the 8th European Conference on Speech Communication and Technology (EUROSPEECH–INTERSPEECH)*. Geneva, Switzerland, 2003, pp. 853–856.

[9] B. Chen, Q. Zhu, and N. Morgan. "Learning Long-Term Temporal Features in LVCSR Using Neural Networks". In: *Proceedings of the 8th International Conference on Spoken Language Processing (INTERSPEECH–ICSLP)*. Jeju, Republic of Korea, 2004, pp. 612–615.

[10] A. Cichocki, S. Cruces, and S. Amari. "Generalized Alpha-Beta Divergences and Their Application to Robust Nonnegative Matrix Factorization". In: *Entropy* 13 (2011), pp. 134–170.

[11] A. Cichocki, R. Zdunek, and S. Amari. "Csiszár's Divergences for Non-Negative Matrix Factorization: Family of New Algorithms". In: *Proceedings of the 6th International Conference on Independent Component Analysis and Blind Source Separation (ICA)*. Charleston, SC, USA, 2006, pp. 32–39.

[12] A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari. *Nonnegative Matrix and Tensor Factorizations*. New York, NY, USA: Wiley, 2009.

[13] M. Cooke, P. Green, and M. Crawford. "Handling missing data in speech recognition". In: *Proceedings of the 3rd International Conference on Spoken Language Processing (ICSLP)*. Yokohama, Japan, 1994, pp. 1555–1558.

[14] M. Cooke, J. Barker, S. Cunningham, and X. Shao. "An Audio-visual Corpus for Speech Perception and Automatic Speech Recognition". In: *Journal of the Acoustical Society of America* 120.5 (2006), pp. 2421–2424.

[15] C. V. Cotton and D. P. W. Ellis. "Spectral vs. Spectro-temporal Features for Acoustic Event Detection". In: *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. New Paltz, NY, USA, 2011, pp. 69–72.

[16] M. Delcroix, K. Kinoshita, T. Nakatani, S. Araki, A. Ogawa, T. Hori, S. Watanabe, M. Fujimoto, T. Yoshioka, T. Oba, Y. Kubo, M. Souden, S. Hahm, and A. Nakamura. "Speech Recognition in the Presence of Highly Non-stationary Noise Based on Spatial, Spectral and Temporal Speech / Noise Modeling Combined with Dynamic Variance Adaptation". In: *Proceedings of the 1st International Workshop on Machine Listening in Multisource Environments (CHiME)*. Florence, Italy, 2011, pp. 12–17.

[17] K. Demuynck, D. Seppi, D. Van Compernolle, P. Nguyen, and G. Zweig. "Integrating Meta-Information into Exemplar-Based Speech Recognition with Segmental Conditional Random Fields". In: *Proceedings of the 36th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Prague, Czech Republic, 2011, pp. 5048–5051.

[18] P. Denes and M. V. Mathews. "Spoken Digit Recognition Using Time-Frequency Pattern Matching". In: *Journal of the Acoustical Society of America* 32 (1960), pp. 1450–1455.

[19] L. Deng and X. Li. "Machine Learning Paradigms for Speech Recognition: An Overview". In: *IEEE Transactions on Audio, Speech, and Language Processing* 21.5 (2013), pp. 1060–1089.

[20] S.-W. Deng and J.-Q. Han. "Statistical voice activity detection based on sparse representation over learned dictionary". In: *Digital Signal Processing* 23.4 (2013), pp. 1228–1232.

[21] M. De Wachter, M. Matton, K. Demuynck, P. Wambacq, R. Cools, and D. Van Compernolle. "Template-based Continuous Speech Recognition". In: *IEEE Transactions on Audio, Speech, and Language Processing* 15 (2007), pp. 1377–1390.

[22] C. Ding, T. Li, and M. I. Jordan. "Convex and semi-nonnegative matrix factorizations". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32.1 (2010), pp. 45–55.

[23] C. Ding, T. Li, W. Peng, and H. Park. "Orthogonal Nonnegative Matrix Tri-factorizations for Clustering". In: *Proceedings of the SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. Philadelphia, PA, USA, 2006, pp. 126–135.

[24] L. Di Persia, D. Milone, H. L. Rufiner, and M. Yanagida. "Perceptual evaluation of blind source separation for robust speech recognition". In: *Signal Processing* 88.10 (2008), pp. 2578–2583.

[25] J. Driesen, J. F. Gemmeke, and H. Van hamme. "Data-driven Speech Representations for NMF-based Word Learning". In: *Proceedings of the Workshop on Statistical and Perceptual Audition with the Speech Communication with Adaptive Learning consortium (SAPA-SCALE)*. Portland, Oregon, USA, 2012, pp. 98–103.

[26] M. Elad. *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. Berlin, Germany: Springer, 2010.

[27] T. M. Elliott and F. E. Frédéric. "The Modulation Transfer Function for Speech Intelligibility". In: *PLoS Computational Biology* 5.3 (2009), e1000302.

[28] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann. “Subjective and Objective Quality Assessment of Audio Source Separation”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 19.7 (2011), pp. 2046–2057.

[29] G. Evermann and P. C. Woodland. “Posterior Probability Decoding, Confidence Estimation, and System Combination”. In: *Proceedings of NIST Speech Transcription Workshop*. College Park, MD, USA, 2000.

[30] C. Févotte, N. Bertin, and J.-L. Durrieu. “Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis”. In: *Neural Computation* 21.3 (2009), pp. 793–830.

[31] C. Févotte and J. Idier. “Algorithms for nonnegative matrix factorization with the β-divergence”. In: *Neural Computation* 23.9 (2011), pp. 2421–2456.

[32] J. G. Fiscus. “A Post-Processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER)”. In: *Proceedings of IEEE Automatic Speech Recognition and Understanding workshop (ASRU)*. Santa Barbara, CA, USA, 1997, pp. 347–354.

[33] D. FitzGerald, M. Cranitch, and E. Coyle. “Non-negative tensor factorisation for sound source separation”. In: *Proceedings of Irish Signals and Systems Conference (ISSC)*. Dublin, Ireland, 2005, pp. 8–12.

[34] J. T. Geiger, R. Vipperla, N. Evans, B. Schuller, and G. Rigoll. “Speech Overlap Detection Using Convolutive Non-Negative Sparse Coding: New Improvements and Insights”. In: *Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*. Bucharest, Romania, 2012, pp. 340–344.

[35] J. T. Geiger, F. Weninger, A. Hurmalainen, J. F. Gemmeke, M. Wöllmer, B. Schuller, G. Rigoll, and T. Virtanen. “The TUM+TUT+KUL Approach to the CHiME Challenge 2013: Multi-Stream ASR Exploiting BLSTM Networks and Sparse NMF”. In: *Proceedings of the 2nd International Workshop on Machine Listening in Multisource Environments (CHiME)*. Vancouver, BC, Canada, 2013, pp. 25–30.

[36] J. T. Geiger, F. Weninger, J. F. Gemmeke, M. Wöllmer, B. Schuller, and G. Rigoll. “Memory-Enhanced Neural Networks and NMF for Robust ASR”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 22.6 (2014), pp. 1037–1046.

[37] J. Gemmeke and B. Cranen. “Noise Robust Digit Recognition Using Sparse Representations”. In: *Proceedings of ISCA ITRW “Speech Analysis and Processing for Knowledge Discovery”*. Aalborg, Denmark, 2008.

[38] J. F. Gemmeke, B. Cranen, and U. Remes. “Sparse imputation for large vocabulary noise robust ASR”. In: *Computer Speech & Language* 25.2 (2011), pp. 462–479.

[39] J. F. Gemmeke and U. Remes. “Missing-Data Techniques: Feature Reconstruction”. In: *Techniques for Noise Robustness in Automatic Speech Recognition*. Ed. by T. Virtanen, R. Singh, and B. Raj. New York, NY, USA: Wiley, 2013.

[40] J. F. Gemmeke and H. Van hamme. “Advances in Noise Robust Digit Recognition using Hybrid Exemplar-Based Techniques”. In: *Proceedings of the 13th Annual Conference of International Speech Communication Association (INTERSPEECH)*. Portland, OR, USA, 2012, pp. 2134–2137.

[41] J. F. Gemmeke and T. Virtanen. “Artificial and Online Acquired Noise Dictionaries for Noise Robust ASR”. In: *Proceedings of the 11th Annual Conference of International Speech Communication Association (INTERSPEECH)*. Makuhari, Japan, 2010, pp. 2082–2085.

[42] J. F. Gemmeke and T. Virtanen. “Noise Robust Exemplar-Based Connected Digit Recognition”. In: *Proceedings of the 35th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Dallas, TX, USA, 2010, pp. 4546–4549.

[43] J. F. Gemmeke, T. Virtanen, and K. Demuynck. “Exemplar-based Joint Channel and Noise Compensation”. In: *Proceedings of the 38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Vancouver, BC, Canada, 2013, pp. 868–872.

[44] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen. “Exemplar-based Sparse Representations for Noise Robust Automatic Speech Recognition”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 19.7 (2011), pp. 2067–2080.

[45] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen. “Exemplar-based Speech Enhancement and Its Application to Noise-robust Automatic Speech Recognition”. In: *Proceedings of the 1st International Workshop on Machine Listening in Multisource Environments (CHiME)*. Florence, Italy, 2011, pp. 53–57.

[46] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen. “HMM-Regularization for NMF-Based Noise Robust ASR”. In: *Proceedings of the 2nd International Workshop on Machine Listening in Multisource Environments (CHiME)*. Vancouver, BC, Canada, 2013, pp. 47–52.

[47] J. F. Gemmeke, L. ten Bosch, L. Boves, and B. Cranen. “Using Sparse Representations for Exemplar Based Continuous Digit Recognition”. In: *Proceedings of the 17th European Signal Processing Conference (EUSIPCO)*. Glasgow, Scotland, UK, 2009, pp. 1755–1759.

[48] J. F. Gemmeke, A. Hurmalainen, T. Virtanen, and Y. Sun. “Toward a Practical Implementation of Exemplar-Based Noise Robust ASR”. In: *Proceedings of the 19th European Signal Processing Conference (EUSIPCO)*. Barcelona, Spain, 2011, pp. 1490–1494.

[49] J. F. Gemmeke, L. Vuegen, P. Karsmakers, B. Vanrumste, and H. Van hamme. “An Exemplar-based NMF Approach to Audio Event Detection”. In: *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. New Paltz, NY, USA, 2013.

[50] Y. Gong. “Speech recognition in noisy environments: A survey”. In: *Speech Communication* 16.3 (1995), pp. 261–291.

[51] E. M. Grais. “Incorporating Prior Information in Nonnegative Matrix Factorization for Audio Source Separation”. PhD thesis. Sabancı University, 2013.

[52] E. M. Grais and H. Erdogan. “Single channel speech music separation using nonnegative matrix factorization with sliding windows and spectral masks”. In: *Proceedings of the 12th Annual Conference of International Speech Communication Association (INTERSPEECH)*. Florence, Italy, 2011, pp. 1773–1776.

[53] E. M. Grais and H. Erdogan. “Hidden Markov Models as Priors for Regularized Nonnegative Matrix Factorization in Singlechannel Source Separation”. In: *Proceedings of the 13th Annual Conference of International Speech Communication Association (INTERSPEECH)*. Portland, OR, USA, 2012, pp. 1536–1539.

[54] E. M. Grais and H. Erdogan. “Spectro-Temporal Post-Smoothing in NMF Based Single-Channel Source Separation”. In: *Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*. Bucharest, Romania, 2012, pp. 584–588.

[55] E. M. Grais and H. Erdogan. “Regularized nonnegative matrix factorization using Gaussian mixture priors for supervised single channel source separation”. In: *Computer Speech & Language* 27.3 (2013), pp. 746–762.

[56] E. M. Grais and H. Erdogan. “Spectro-temporal post-enhancement using MMSE estimation in NMF based single-channel source separation”. In: *Proceedings of the 14th Annual Conference of International Speech Communication Association (INTERSPEECH)*. Lyon, France, 2013, pp. 3279–3283.

[57] D. W. Griffin and J. S. Lim. “Signal Estimation from Modified Short-Time Fourier Transform”. In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 32.2 (1984), pp. 236–243.

[58] J. H. L. Hansen and B. L. Pellom. “An Effective Quality Evaluation Protocol for Speech Enhancement Algorithms”. In: *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP)*. Sydney, Australia, 1998, pp. 2819–2822.

[59] Y. Han, J. de Veth, and L. Boves. “Trajectory Clustering for Solving the Trajectory Folding Problem in Automatic Speech Recognition”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 15.4 (2007), pp. 1425–1434.

[60] G. Heigold, P. Nguyen, M. Weintraub, and V. Vanhoucke. “Investigations on Exemplar-Based Features for Speech Recognition Towards Thousands of Hours of Unsupervised, Noisy Data”. In: *Proceedings of the 37th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Kyoto, Japan, 2012, pp. 4437–4440.

[61] T. Heittola, A. Mesaros, T. Virtanen, and A. Eronen. “Sound Event Detection in Multisource Environments Using Source Separation”. In: *Proceedings of the 1st International Workshop on Machine Listening in Multisource Environments (CHiME)*. Florence, Italy, 2011, pp. 36–40.

[62] M. Helén and T. Virtanen. “Separation of Drums from Polyphonic Music Using Non-negative Matrix Factorization and Support Vector Machine”. In: *Proceedings of the 13th European Signal Processing Conference (EUSIPCO)*. Antalya, Turkey, 2005.

[63] H. Hermansky. “Perceptual linear predictive (PLP) analysis of speech”. In: *Journal of the Acoustical Society of America* 87.4 (1990), pp. 1738–1752.

[64] H. Hermansky, D. P. W. Ellis, and S. Sharma. “Tandem Connectionist Feature Extraction for Conventional HMM Systems”. In: *Proceedings of the 25th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Istanbul, Turkey, 2000, pp. III 1635–1638.

[65] H. Hermansky and S. Sharma. “TRAPs – Classifiers of Temporal Patterns”. In: *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP)*. Sydney, Australia, 1998, pp. 1003–1006.

[66] M. B. Herscher and R. B. Cox. “An Adaptive Isolated-Word Speech Recognition System”. In: *Proceedings of IEEE Conference on Speech Communication and Processing*. Newton, MA, USA, 1972, pp. 89–92.

[67] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury. “Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups”. In: *IEEE Signal Processing Magazine* 29.6 (2012), pp. 82–97.

[68] H.-G. Hirsch and D. Pearce. “The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions”. In: *Proceedings of ASR2000 — Automatic Speech Recognition: Challenges for the new Millenium, ISCA Tutorial and Research Workshop (ITRW)*. Paris, France, 2000, pp. 181–188.

[69] P. O. Hoyer. “Non-Negative Sparse Coding”. In: *Proceedings of IEEE Workshop on Neural Networks for Signal Processing (NNSP)*. Martigny, Switzerland, 2002, pp. 557–565.

[70] A. Hurmalainen and T. Virtanen. “Learning State Labels for Sparse Classification of Speech with Matrix Deconvolution”. In: *Proceedings of IEEE Automatic Speech Recognition and Understanding workshop (ASRU)*. Olomouc, Czech Republic, 2013, pp. 168–173.

[71] Y. Hu and P. C. Loizou. “Subjective comparison and evaluation of speech enhancement algorithms”. In: *Speech Communication* 49.7 (2007), pp. 588–601.

[72] A. Jansen and K. Church. “Towards Unsupervised Training of Speaker Independent Acoustic Models”. In: *Proceedings of the 12th Annual Conference of International Speech Communication Association (INTERSPEECH)*. Florence, Italy, 2011, pp. 1693–1696.

[73] K. M. Jeon, H. K. Kim, S. J. Lee, and Y. K. Lee. “Non-negative Matrix Factorization Based Adaptive Noise Sensing over Wireless Sensor Networks”. In: *International Journal of Distributed Sensor Networks* 2014 Article ID 640915 (2014), 9 pages.

[74] C. Joder and B. Schuller. “Exploring Nonnegative Matrix Factorization for Audio Classification: Application to Speaker Recognition”. In: *Proceedings of ITG Conference on Speech Communication*. Braunschweig, Germany, 2012.

[75] C. Joder, F. Weninger, F. Eyben, D. Virette, and B. Schuller. “Real-Time Speech Separation by Semi-supervised Nonnegative Matrix Factorization”. In: *Proceedings of the 10th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*. Tel Aviv, Israel, 2012.

[76] C. Joder, F. Weninger, D. Virette, and B. Schuller. “A Comparative Study on Sparsity Penalties for NMF-based Speech Separation: Beyond LP-Norms”. In: *Proceedings of the 38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Vancouver, BC, Canada, 2013, pp. 858–862.

[77] C. Joder, F. Weninger, D. Virette, and B. Schuller. “Integrating Noise Estimation and Factorization-Based Speech Separation: A Novel Hybrid Approach”. In: *Proceedings of the 38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Vancouver, BC, Canada, 2013, pp. 131–135.

[78] H. Kallasjoki, J. F. Gemmeke, and K. J. Palomäki. “Estimating Uncertainty to Improve Exemplar-Based Feature Enhancement for Noise Robust Speech Recognition”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 22.2 (2014), pp. 368–380.

[79] H. Kallasjoki, U. Remes, J. F. Gemmeke, T. Virtanen, and K. J. Palomäki. “Uncertainty Measures for Improving Exemplar-Based Source Separation”. In: *Proceedings of the 12th Annual Conference of International Speech Communication Association (INTERSPEECH)*. Florence, Italy, 2011, pp. 469–472.

[80] S. Keronen, H. Kallasjoki, U. Remes, G. J. Brown, J. F. Gemmeke, and K. J. Palomäki. “Mask estimation and imputation methods for missing data speech recognition in a multisource reverberant environment”. In: *Computer Speech & Language* 27.3 (2013), pp. 798–819.

[81] M. Kim, J. Yoo, K. Kang, and S. Choi. “Nonnegative matrix partial co-factorization for spectral and temporal drum source separation”. In: *IEEE Journal of Selected Topics in Signal Processing* 5.6 (2011), pp. 1192–1204.

[82] K. Kinoshita, M. Souden, M. Delcroix, and T. Nakatani. “Single Channel Dereverberation Using Example-Based Speech Enhancement with Uncertainty Decoding Technique”. In: *Proceedings of the 12th Annual Conference of International Speech Communication Association (INTERSPEECH)*. Florence, Italy, 2011, pp. 197–200.

[83] M. Kleinschmidt. “Methods for Capturing Spectro-temporal Modulations in ASR”. In: *Acta Acustica United with Acustica* 88.3 (2002), pp. 416–422.

[84] M. Kleinschmidt and D. Gelbart. “Improving Word Accuracy with Gabor Feature Extraction”. In: *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP–INTERSPEECH)*. Denver, Colorado, USA, 2002, pp. 25–28.

[85] D. Kolossa and R. Haeb-Umbach, eds. *Robust Speech Recognition of Uncertain or Missing Data*. Berlin, Germany: Springer, 2011.

[86] J. M. K. Kua, E. Ambikairajah, J. Epps, and R. Togneri. “Speaker Verification Using Sparse Representation Classification”. In: *Proceedings of the 36th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Prague, Czech Republic, 2011, pp. 4548–4551.

[87] P. Langjahr and P. Mowlaee. “Objective Quality Assessment of Target Speaker Separation Performance in Multisource Reverberant Environment”. In: *Proceedings of 4th International Workshop on Perceptual Quality of Systems (PQS)*. Vienna, Austria, 2013, pp. 89–94.

[88] B. Lecouteux, G. Linarès, Y. Estève, and G. Gravier. “Dynamic Combination of Automatic Speech Recognition Systems by Driven Decoding”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 21.6 (2013), pp. 1251–1260.

[89] C. Lee and J. Glass. “A Nonparametric Bayesian Approach to Acoustic Model Discovery”. In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*. Jeju, Republic of Korea, 2012, pp. 40–49.

[90] C.-T. Lee, Y.-H. Yang, and H.-H. Chen. “Multipitch Estimation of Piano Music by Exemplar-Based Sparse Representation”. In: *IEEE Transcations on Multimedia* 14.3 (2012), pp. 608–618.

[91] D. D. Lee and H. S. Seung. “Learning the parts of objects by non-negative matrix factorization”. In: *Nature* 401.6755 (1999), pp. 788–791.

[92] D. D. Lee and H. S. Seung. “Algorithms for Non-negative Matrix Factorization”. In: *Advances in Neural Information Processing Systems 13*. 2001, pp. 556–562.

[93] A. Lefèvre, F. Bach, and C. Févotte. “Itakura-Saito Nonnegative Matrix Factorization with Group Sparsity”. In: *Proceedings of the 36th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Prague, Czech Republic, 2011, pp. 21–24.

[94] A. Lefèvre, F. Bach, and C. Févotte. “Online Algorithms for Nonnegative Matrix Factorization with the Itakura-Saito Divergence”. In: *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. New Paltz, NY, USA, 2011.

[95] R. Leonard. “A Database for Speaker-Independent Digit Recognition”. In: *Proceedings of the 9th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. San Diego, CA, USA, 1984, pp. 328–331.

[96] J. Le Roux, E. Vincent, Y. Mizuno, H. Kameoka, N. Ono, and S. Sagayama. “Consistent Wiener Filtering: Generalized Time-Frequency Masking Respecting Spectrogram Consistency”. In: *Proceedings of the 9th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*. St. Malo, France, 2010, pp. 89–96.

[97] S. Levinson, L. Rabiner, A. Rosenberg, and J. Wilpon. “Interactive clustering techniques for selecting speaker-independent reference templates for isolated word recognition”. In: *IEEE Transactions on Acoustics, Speech and Signal Processing* 27.3 (1979), pp. 134–141.

[98] Y. Liao, C. Lin, and W. Fang. “Minimum Classification Error Based Spectro-Temporal Feature Extraction for Robust Audio Classification”. In: *Proceedings of the 12th Annual Conference of International Speech Communication Association (INTERSPEECH)*. Florence, Italy, 2011, pp. 241–244.

[99] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach. “An Overview of Noise-Robust Automatic Speech Recognition”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 22.4 (2014), pp. 745–777.

[100] R. P. Lippmann. “Review of Neural Networks for Speech Recognition”. In: *Neural Computation* 1.1 (1989), pp. 1–38.

[101] Q. Li. “An Auditory-Based Transform for Audio Signal Processing”. In: *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. New Paltz, NY, USA, 2009, pp. 181–184.

[102] W. Liu, N. Zheng, and X. Lu. “Non-negative Matrix Factorization for Visual Coding”. In: *Proceedings of the 28th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Hong Kong, 2003, pp. III 293–296.

[103] P. C. Loizou. *Speech Enhancement: Theory and Practice, second edition*. Boca Raton, FL, USA: CRC Press, 2013.

[104] Y.-H. Long, L.-R. Dai, E.-Y. Wang, B. Ma, and W. Guo. “Non-negative matrix factorization based discriminative features for speaker verification”. In: *Proceedings of the 7th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. Tainan, Taiwan, 2010, pp. 291–295.

[105] H. Lu, Z. Fu, and X. Shu. “Non-negative and sparse spectral clustering”. In: *Pattern Recognition* 47.1 (2014), pp. 418–426.

[106] K. Mahkonen, A. Hurmalainen, T. Virtanen, and J. Gemmeke. “Mapping Sparse Representation to State Likelihoods in Noise-Robust Automatic Speech Recognition”. In: *Proceedings of the 12th Annual Conference of International Speech Communication Association (INTERSPEECH)*. Florence, Italy, 2011, pp. 465–468.

[107] S. Makino, T.-W. Lee, and H. Sawada, eds. *Blind Speech Separation*. Berlin, Germany: Springer, 2007.

[108] M.-W. Mak and H.-B. Yu. “A study of voice activity detection techniques for NIST speaker recognition evaluations”. In: *Computer Speech & Language* 28.1 (2014), pp. 295–313.

[109] N. Mesgarani and S. Shamma. “Speech Processing with a Cortical Representation of Audio”. In: *Proceedings of the 36th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Prague, Czech Republic, 2011, pp. 5872–5875.

[110] N. Mesgarani, S. V. David, J. B. Fritz, and S. A. Shamma. “Phoneme representation and classification in primary auditory cortex”. In: *Journal of the Acoustical Society of America* 132.2 (2008), pp. 899–909.

[111] B. T. Meyer and B. Kollmeier. “Optimization and evaluation of Gabor feature sets for ASR”. In: *Proceedings of the 9th Annual Conference of International Speech Communication Association (INTERSPEECH)*. Brisbane, Australia, 2008, pp. 906–909.

[112] B. T. Meyer, S. V. Ravuri, M. R. Schadler, and N. Morgan. “Comparing Different Flavors of Spectro-Temporal Features for ASR”. In: *Proceedings of the 12th Annual Conference of International Speech Communication Association (INTERSPEECH)*. Florence, Italy, 2011, pp. 1269–1272.

[113] A. Mohamed, G. Dahl, and G. Hinton. “Acoustic Modeling using Deep Belief Networks”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 20.1 (2012), pp. 14–22.

[114] N. Mohammadiha, P. Smaragdis, and A. Leijon. “Prediction Based Filtering and Smoothing to Exploit Temporal Dependencies in NMF”. In: *Proceedings of the 38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Vancouver, BC, Canada, 2013, pp. 873–877.

[115] N. Mohammadiha, P. Smaragdis, and A. Leijon. “Supervised and Unsupervised Speech Enhancement Using Nonnegative Matrix Factorization”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 21.10 (2013), pp. 2140–2151.

[116] N. Morgan and H. Bourlard. “Continuous Speech Recognition Using Multilayer Perceptrons with Hidden Markov Models”. In: *Proceedings of the 15th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Albuquerque, NM, USA, 1990, pp. 413–416.

[117] P. Mowlaee, R. Saeidi, M. G. Christensen, and R. Martin. “Subjective and Objective Quality Assessment of Single-Channel Speech Separation Algorithms”. In: *Proceedings of the 37th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Kyoto, Japan, 2012, pp. 69–72.

[118] G. J. Mysore and P. Smaragdis. “A Non-negative Approach to Semi-Supervised Separation of Speech from Noise with the Use of Temporal Dynamics”. In: *Proceedings of the 36th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Prague, Czech Republic, 2011, pp. 17–20.

[119] G. Mysore, P. Smaragdis, and B. Raj. “Non-negative Hidden Markov Modeling of Audio with Application to Source Separation”. In: *Proceedings of the 9th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*. St. Malo, France, 2010.

[120] I. Naseem, R. Togneri, and M. Bennamoun. “Sparse Representation for Speaker Identification”. In: *Proceedings of the International Conference on Pattern Recognition (ICPR)*. Istanbul, Turkey, 2010, pp. 4460–4463.

[121] T. Oates. “PERUSE: An Unsupervised Algorithm for Finding Recurrent Patterns in Time Series”. In: *Proceedings of IEEE International Conference on Data Mining (ICDM)*. Maebashi City, Japan, 2002, pp. 330–337.

[122] P. D. O’Grady and B. A. Pearlmutter. “Discovering Convolutive Speech Phones using Sparseness and Non-Negativity Constraints”. In: *Proceedings of the 7th International Conference on Independent Component Analysis and Signal Separation (ICA)*. London, UK, 2007, pp. 520–527.

[123] A. Ozerov and C. Févotte. “Multichannel Nonnegative Matrix Factorization in Convolutive Mixtures for Audio Source Separation”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 18.3 (2010), pp. 550–563.

[124] A. Ozerov, C. Févotte, and M. Charbit. “Factorial Scaled Hidden Markov Model for Polyphonic Audio Representation and Source Separation”. In: *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA).* New Paltz, NY, USA, 2009, pp. 121–124.

[125] A. Ozerov, E. Vincent, and F. Bimbot. “A general flexible framework for the handling of prior information in audio source separation”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 20.4 (2012), pp. 1118–1133.

[126] Y. Panagakis, C. Kotropoulos, and G. R. Arce. “Music Genre Classification via Sparse Representations of Auditory Temporal Modulations”. In: *Proceedings of the 17th European Signal Processing Conference (EUSIPCO).* Glasgow, Scotland, UK, 2009, pp. 1–5.

[127] R. M. Parry and I. A. Essa. “Estimating the Spatial Position of Spectral Components in Audio”. In: *Proceedings of the 6th International Conference on Independent Component Analysis and Blind Source Separation (ICA).* Charleston, SC, USA, 2006, pp. 666–673.

[128] B. N. Pasley, S. V. David, N. Mesgarani, A. Flinker, S. A. Shamma, N. E. Crone, R. T. Knight, and E. F. Chang. “Reconstructing Speech from Human Auditory Cortex”. In: *PLoS Biology* 10.1 (2012), e1001251.

[129] Y. Pereiro Estevan, V. Wan, and O. Scharenborg. “Finding Maximum Margin Segments in Speech”. In: *Proceedings of the 32nd International Conference on Acoustics, Speech, and Signal Processing (ICASSP).* Honolulu, HI, USA, 2007, pp. IV 937–940.

[130] J. O. Pickles. *An Introduction to the Physiology of Hearing, 4th edition.* Leiden, Netherlands: Brill, 2013.

[131] Y. Qiao, N. Shimomura, and N. Minematsu. “Unsupervised Optimal Phoneme Segmentation: Objectives, Algorithm and Comparisons”. In: *Proceedings of the 33rd International Conference on Acoustics, Speech, and Signal Processing (ICASSP).* Las Vegas, NV, USA, 2008, pp. 3989–3992.

[132] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements. *Objective Measures of Speech Quality.* NJ, USA: Prentice Hall, 1988.

[133] L. R. Rabiner and B.-H. Juang. *Fundamentals of speech recognition*. Englewood Cliffs, NJ, USA: Prentice Hall, 1993.

[134] L. R. Rabiner and J. G. Wilpon. “Speaker-Independent Isolated Word Recognition for a Moderate Size (54 Word) Vocabulary”. In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 6 (1979), pp. 583–587.

[135] B. Raj, R. Singh, and T. Virtanen. “Phoneme-dependent NMF for speech enhancement in monaural mixtures”. In: *Proceedings of the 12th Annual Conference of International Speech Communication Association (INTERSPEECH)*. Florence, Italy, 2011, pp. 1217–1220.

[136] B. Raj and R. M. Stern. “Missing-feature approaches in speech recognition”. In: *IEEE Signal Processing Magazine* 22.5 (2005), pp. 101–116.

[137] B. Raj, T. Virtanen, S. Chaudhure, and R. Singh. “Non-negative Matrix Factorization Based Compensation of Music for Automatic Speech Recognition”. In: *Proceedings of the 11th Annual Conference of International Speech Communication Association (INTERSPEECH)*. Makuhari, Japan, 2010, pp. 717–720.

[138] O. Räsänen. “A computational model of word segmentation from continuous speech using transitional probabilities of atomic acoustic events”. In: *Cognition* 120.2 (2011), pp. 149–176.

[139] O. Räsänen and J. Driesen. “A comparison and combination of segmental and fixed-frame signal representations in NMF-based word recognition”. In: *Proceedings of the 17th Nordic Conference of Computational Linguistics (NoDaLiDa)*. Odense, Denmark, 2009, pp. 255–262.

[140] O. Räsänen and U.K. Laine. “A method for noise-robust context-aware pattern discovery and recognition from categorical sequences”. In: *Pattern Recognition* 45.1 (2012), pp. 606–616.

[141] S. Renals, N. Morgan, H. Boulard, M. Cohen, and H. Franco. “Connectionist Probability Estimators in HMM Speech Recognition”. In: *IEEE Transactions on Speech and Audio Processing* 2.1 (1994), pp. 161–174.

[142] R. Rui and C.-C. Bao. “Projective Non-negative Matrix Factorization with Bregman Divergence for Musical Instrument Classification”. In: *Proceedings of IEEE International Conference on Signal Processing, Communication and Computing (ICSPCC)*. Hong Kong, 2012, pp. 415–418.

[143] R. Saeidi, A. Hurmalainen, T. Virtanen, and D. A. van Leeuwen. “Exemplar-based Sparse Representation and Sparse Discrimination for Noise Robust Speaker Identification”. In: *Odyssey speaker and language recognition workshop*. Singapore, 2012.

[144] M. Sahidullah and G. Saha. “Comparison of Speech Activity Detection Techniques for Speaker Recognition”. In: *CoRR* arXiv: 1210.0297 (2012).

[145] T. N. Sainath, B. Ramabhadran, M. Picheny, D. Nahamoo, and D. Kanevsky. “Exemplar-Based Sparse Representation Features: From TIMIT to LVCSR”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 19.8 (2011), pp. 2598–2613.

[146] T. N. Sainath, B. Ramabhadran, D. Nahamoo, D. Kanevsky, D. Van Compernolle, K. Demuynck, J. F. Gemmeke, J. R. Bellegarda, and S. Sundaram. “Exemplar-Based Processing for Speech Recognition: An Overview”. In: *IEEE Signal Processing Magazine* 29.6 (2012), pp. 98–113.

[147] A. Sankar. “Bayesian Model Combination (BAYCOM) for Improved Recognition”. In: *Proceedings of the 30th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Philadelphia, PA, USA, 2005, pp. 845–848.

[148] H. Sawada, H. Kameoka, S. Araki, and N. Ueda. “New Formulations and Efficient Algorithms for Multichannel NMF”. In: *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. New Paltz, NY, USA, 2011, pp. 132–156.

[149] H. Sawada, H. Kameoka, S. Araki, and N. Ueda. “Multichannel Extensions of Non-Negative Matrix Factorization With Complex-Valued Data”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 21.5 (2013), pp. 971–982.

[150] O. Scharenborg. “Reaching over the gap: A review of efforts to link human and automatic speech recognition research”. In: *Speech Communication* 49.5 (2007), pp. 336–347.

[151] O. Scharenborg, V. Wan, and M. Ernestus. “Unsupervised speech segmentation: An analysis of the hypothesized phone boundaries”. In: *Journal of the Acoustical Society of America* 127.2 (2010), pp. 1084–1095.

[152] M. N. Schmidt and M. Mørup. “Nonnegative Matrix Factor 2-D Deconvolution for Blind Single Channel Source Separation”. In: *Proceedings of the 6th International Conference on Independent Component Analysis and Blind Source Separation (ICA)*. Charleston, SC, USA, 2006, pp. 700–707.

[153] M. N. Schmidt and R. K. Olsson. “Single-channel Speech Separation using Sparse Non-negative Matrix Factorization”. In: *Proceedings of the 9th International Conference on Spoken Language Processing (INTERSPEECH-ICSLP)*. Pittsburgh, PA, USA, 2006, pp. 2614–2617.

[154] B. Schuller and F. Weninger. “Discrimination of Speech and Non-Linguistic Vocalizations by Non-Negative Matrix Factorization”. In: *Proceedings of the 35th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Dallas, TX, USA, 2010, pp. 5054–5057.

[155] Y. Shao, Z. Jin, D. Wang, and S. Srinivasan. “An Auditory-Based Feature for Robust Speech Recognition”. In: *Proceedings of the 34th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Taipei, Taiwan, 2009, pp. 4625–2628.

[156] M. Slaney, D. Naar, and R. F. Lyon. “Auditory Model Inversion for Sound Separation”. In: *Proceedings of the 19th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Adelaide, Australia, 1994, pp. II 77–80.

[157] P. Smaragdis. “Non-negative Matrix Factor Deconvolution; Extraction of Multiple Sound Sources from Monophonic Inputs”. In: *Proceedings of the 5th International Conference on Independent Component Analysis and Blind Signal Separation (ICA)*. Grenada, Spain, 2004, pp. 494–499.

[158] P. Smaragdis. “Convolutive Speech Bases and their Application to Supervised Speech Separation”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 15.1 (2007), pp. 1–14.

[159] P. Smaragdis and J. C. Brown. “Non-Negative Matrix Factorization for Polyphonic Music Transcription”. In: *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. New Paltz, NY, USA, 2003, pp. 177–180.

[160] P. Smaragdis, C. Févotte, G. J. Mysore, N. Mohammadiha, and M. Hoffman. “Static and Dynamic Source Separation Using Nonnegative Factorizations — A unified view”. In: *IEEE Signal Processing Magazine* 31.3 (2014), pp. 66–75.

[161] P. Sprechmann, A. M. Bronstein, M. M. Bronstein, and G. Sapiro. “Learnable Low Rank Sparse Models for Speech Denoising”. In: *Proceedings of the 38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Vancouver, BC, Canada, 2013, pp. 136–140.

[162] R. M. Stern and N. Morgan. “Features Based on Auditory Physiology and Perception”. In: *Techniques for Noise Robustness in Automatic Speech Recognition*. Ed. by T. Virtanen, R. Singh, and B. Raj. New York, NY, USA: Wiley, 2013.

[163] N. Sturmel and L. Daudet. “Iterative Phase Reconstruction of Wiener Filtered Signals”. In: *Proceedings of the 37th International Conference on Acoustics, Speech, and Signal Processing (ICASSP).* Kyoto, Japan, 2012, pp. 101–104.

[164] S. Sundaram and J. Bellegarda. “Latent Perceptual Mapping with Data-Driven Variable-Length Acoustic Units for Template-Based Speech Recognition”. In: *Proceedings of the 37th International Conference on Acoustics, Speech, and Signal Processing (ICASSP).* Kyoto, Japan, 2012, pp. 4125–4128.

[165] D. L. Sun and G. J. Mysore. “Universal Speech Models for Speaker Independent Single Channel Source Separation”. In: *Proceedings of the 38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP).* Vancouver, BC, Canada, 2013, pp. 141–145.

[166] X. Sun and Y. Zhao. “Integrated exemplar-based template matching and statistical modeling for continuous speech recognition”. In: *EURASIP Journal on Audio, Speech, and Music Processing* 2014.4 (2014), 16 pages.

[167] Y. Sun, J. F. Gemmeke, B. Cranen, L. ten Bosch, and L. Boves. “Using a DBN to Integrate Sparse Classification and GMM-based ASR”. In: *Proceedings of the 11th Annual Conference of International Speech Communication Association (INTERSPEECH).* Makuhari, Japan, 2010, pp. 2098–2101.

[168] Y. Sun, J. F. Gemmeke, B. Cranen, L. ten Bosch, and L. Boves. “Improvements of a Dual-input DBN for Noise Robust ASR”. In: *Proceedings of the 12th Annual Conference of International Speech Communication Association (INTERSPEECH).* Florence, Italy, 2011, pp. 1669–1672.

[169] Y. Sun, M. M. Doss, J. F. Gemmeke, B. Cranen, L. ten Bosch, and L. Boves. “Combination of Sparse Classification and Multilayer Perceptron for Noise-robust ASR”. In: *Proceedings of the 13th Annual Conference of International Speech Communication Association (INTERSPEECH).* Portland, OR, USA, 2012, pp. 310–313.

[170] Y. Sun, B. Cranen, J. F. Gemmeke, L. Boves, L. ten Bosch, and M. M. Doss. “Using Sparse Classification Outputs as Feature Observations for Noise-robust ASR”. In: *Proceedings of the 13th Annual Conference of International Speech Communication Association (INTERSPEECH).* Portland, OR, USA, 2012, pp. 2142–2145.

[171] R. Takashima, T. Takiguchi, and Y. Ariki. “Exemplar-based Voice Conversion in Noisy Environment”. In: *Proceedings of IEEE Spoken Language Technology Workshop (SLT)*. Miami, FL, USA, 2012, pp. 313–317.

[172] Q. Tan and S. Narayanan. “Novel Variations of Group Sparse Regularization Techniques with Applications to Noise Robust Automatic Speech Recognition”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 20.4 (2012), pp. 1337–1346.

[173] L. ten Bosch. “Bridging the gap between human and automatic speech recognition”. In: *Speech Communication* 49.5 (2007), pp. 331–335.

[174] C. Thurau, K. Kersting, and C. Bauckhage. “Convex Non-negative Matrix Factorization in the Wild”. In: *Proceedings of International Conference on Data Mining (ICDM)*. Miami, FL, USA, 2009, pp. 523–532.

[175] D. T. Tran, E. Vincent, D. Jouvet, and K. Adiloğlu. “Using Full-Rank Spatial Covariance Models for Noise-Robust ASR”. In: *Proceedings of the 2nd International Workshop on Machine Listening in Multisource Environments (CHiME)*. Vancouver, BC, Canada, 2013, pp. 31–32.

[176] Z. Tüske, R. Schlüter, H. Ney, and M. Sundermeyer. “Context-Dependent MLPs for LVCSR: TANDEM, Hybrid or Both?” In: *Proceedings of the 13th Annual Conference of International Speech Communication Association (INTERSPEECH)*. Portland, OR, USA, 2012, pp. 18–21.

[177] C. Tzagkarakis and A. Mouchtaris. “Sparsity Based Noise Robust Speaker Identification Using a Discriminative Dictionary Learning Approach”. In: *Proceedings of the 21st European Signal Processing Conference (EUSIPCO)*. Marrakech, Morocco, 2013.

[178] H. Van hamme. “HAC-models: a Novel Approach to Continuous Speech Recognition”. In: *Proceedings of the 9th Annual Conference of International Speech Communication Association (INTERSPEECH)*. Brisbane, Australia, 2008, pp. 2554–2557.

[179] H. Van hamme. “A Diagonalized Newton Algorithm for Non-negative Sparse Coding”. In: *Proceedings of the 38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Vancouver, BC, Canada, 2013, pp. 7299–7303.

[180] M. Van Segbroeck and H. Van hamme. “Unsupervised Learning of Time-Frequency Patches as a Noise-robust Representation of Speech”. In: *Speech Communication* 51.11 (2009), pp. 1124–1138.

[181] O. Viikki and K. Laurila. “Cepstral domain segmental feature vector normalization for noise robust speech recognition”. In: *Speech Communication* 25.1 (1998), pp. 133–147.

[182] E. Vincent, R. Gribonval, and C. Févotte. “Performance measurement in blind audio source separation”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 14.4 (2006), pp. 1462–1469.

[183] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni. “The Second CHiME Speech Separation and Recognition Challenge: an Overview of Challenge Systems and Outcomes”. In: *Proceedings of IEEE Automatic Speech Recognition and Understanding workshop (ASRU).* Olomouc, Czech Republic, 2013, pp. 162–167.

[184] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni. “The Second ‘CHiME’ Speech Separation and Recognition Challenge: Datasets, Tasks and Baselines”. In: *Proceedings of the 38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP).* Vancouver, BC, Canada, 2013, pp. 126–130.

[185] R. Vipperla, S. Bozonnet, D. Wang, and N. Evans. “Robust Speech Recognition in Multi-Source Noise Environments using Convolutive Non-Negative Matrix Factorization”. In: *Proceedings of the 1st International Workshop on Machine Listening in Multisource Environments (CHiME).* Florence, Italy, 2011, pp. 74–79.

[186] R. Vipperla, J. Geiger, S. Bozonnet, D. Wang, N. Evans, B. Schuller, and G. Rigoll. “Speech Overlap Detection and Attribution Using Convolutive Non-Negative Sparse Coding”. In: *Proceedings of the 37th International Conference on Acoustics, Speech, and Signal Processing (ICASSP).* Kyoto, Japan, 2012, pp. 4181–4184.

[187] T. Virtanen. “Sound Source Separation Using Sparse Coding with Temporal Continuity Objective”. In: *Proceedings of International Computer Music Conference (ICMC).* Singapore, 2003.

[188] T. Virtanen. “Separation of Sound Sources by Convolutive Sparse Coding”. In: *Proceedings of ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing (SAPA).* Jeju, Republic of Korea, 2004.

[189] T. Virtanen. “Monaural Sound Source Separation by Non-Negative Matrix Factorization with Temporal Continuity and Sparseness Criteria”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 15.3 (2007), pp. 1066–1074.

[190] T. Virtanen, J. F. Gemmeke, and A. Hurmalainen. “State-based Labelling for a Sparse Representation of Speech and Its Application to Robust Speech Recognition”. In: *Proceedings of the 11th Annual Conference*

of International Speech Communication Association (INTERSPEECH). Makuhari, Japan, 2010, pp. 893–896.

[191] T. Virtanen, J. F. Gemmeke, and B. Raj. “Active-Set Newton Algorithm for Overcomplete Non-Negative Representations of Audio”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 21.11 (2013), pp. 2277–2289.

[192] T. Virtanen and A. Klapuri. “Analysis of polyphonic audio using source-filter model and non-negative matrix factorization”. In: *Proceedings of Advances in Models for Acoustic Processing, Neural Information Processing Systems Workshop (NIPS).* Whistler, BC, Canada, 2006.

[193] T. Virtanen, R. Singh, and B. Raj, eds. *Techniques for Noise Robustness in Automatic Speech Recognition.* New York, NY, USA: Wiley, 2012.

[194] T. Virtanen, J. F. Gemmeke, B. Raj, and P. Smaragdis. “Compositional models for audio processing”. In: *IEEE Signal Processing Magazine* (2014).

[195] B. Wang and M. D. Plumbley. “Musical Audio Stream Separation by Non-negative Matrix Factorization”. In: *Proceedings of the UK Digital Music Research Network (DMRN) Summer Conference.* Glasgow, Scotland, 2005.

[196] D. Wang and J. Tejedor. “Heterogeneous Convolutive Non-Negative Sparse Coding”. In: *Proceedings of the 13th Annual Conference of International Speech Communication Association (INTERSPEECH).* Portland, OR, USA, 2012, pp. 2150–2153.

[197] D. Wang, R. Vipperla, and N. Evans. “Online Pattern Learning for Non-Negative Convolutive Sparse Coding”. In: *Proceedings of the 12th Annual Conference of International Speech Communication Association (INTERSPEECH).* Florence, Italy, 2011, pp. 65–68.

[198] W. Wang. “Convolutive Non-Negative Sparse Coding”. In: *Proceedings of International Joint Conference on Neural Networks (IJCNN).* Hong Kong, 2008, pp. 3681–3684.

[199] F. Weninger, J. Feliu, and B. Schuller. “Supervised and Semi-supervised Suppression of Background Music in Monaural Speech Recordings”. In: *Proceedings of the 37th International Conference on Acoustics, Speech, and Signal Processing (ICASSP).* Kyoto, Japan, 2012, pp. 61–64.

[200] F. Weninger, A. Lehmann, and B. Schuller. “OpenBliSSART: Design and Evaluation of a Research Toolkit for Blind Source Separation in Audio Recognition Tasks”. In: *Proceedings of the 36th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Prague, Czech Republic, 2011, pp. 1625–1628.

[201] F. Weninger and B. Schuller. “Optimization and Parallelization of Monaural Source Separation Algorithms in the openBliSSART Toolkit”. In: *Journal of Signal Processing Systems* 69.3 (2012), pp. 267–277.

[202] F. Weninger, M. Wöllmer, and B. Schuller. “Sparse, Hierarchical and Semi-Supervised Base Learning for Monaural Enhancement of Conversational Speech”. In: *Proceedings of ITG Conference on Speech Communication*. Braunschweig, Germany, 2012.

[203] F. Weninger, J. Geiger, M. Wöllmer, B. Schuller, and G. Rigoll. “The Munich 2011 CHiME Challenge Contribution: NMF-BLSTM Speech Enhancement and Recognition for Reverberated Multisource Environments”. In: *Proceedings of the 1st International Workshop on Machine Listening in Multisource Environments (CHiME)*. Florence, Italy, 2011, pp. 24–29.

[204] F. Weninger, M. Wöllmer, J. Geiger, B. Schuller, J. F. Gemmeke, A. Hurmalainen, T. Virtanen, and G. Rigoll. “Non-negative Matrix Factorization for Highly Noise-robust ASR: To Enhance or to Recognize?” In: *Proceedings of the 37th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Kyoto, Japan, 2012, pp. 4681–4684.

[205] K. W. Wilson, B. Raj, and P. Smaragdis. “Regularized Non-Negative Matrix Factorization with Temporal Dependencies for Speech Denoising”. In: *Proceedings of the 9th Annual Conference of International Speech Communication Association (INTERSPEECH)*. Brisbane, Australia, 2008, pp. 411–414.

[206] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran. “Speech Denoising Using Nonnegative Matrix Factorization with Priors”. In: *Proceedings of the 33rd International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Las Vegas, NV, USA, 2008, pp. 4029–4032.

[207] M. Wölfel and J. McDonough. *Distant Speech Recognition*. New York, NY, USA: Wiley, 2009.

[208] M. Wöllmer, F. Weninger, J. Geiger, B. Schuller, and G. Rigoll. “Noise robust ASR in reverberated multisource environments applying convolutive NMF and Long Short-Term Memory”. In: *Computer Speech & Language* 27.3 (2013), pp. 780–797.

[209] Q. Wu, L.-Q. Zhang, and G.-C. Shi. “Robust Feature Extraction for Speaker Recognition Based on Constrained Nonnegative Tensor Factorization”. In: *Journal of Computer Science and Technology* 25.4 (2010), pp. 745–754.

[210] Z. Wu, T. Virtanen, T. Kinnunen, E. S. Chng, and H. Li. “Exemplar-based Voice Conversion using Non-negative Spectrogram Deconvolution”. In: *Proceedings of the 8th ISCA Speech Synthesis Workshop*. Barcelona, Spain, 2013, pp. 201–206.

[211] E. Yılmaz, J. F. Gemmeke, and H. Van hamme. “Noise-robust Speech Recognition with Exemplar-based Sparse Representations Using Alpha-Beta Divergence”. In: *Proceedings of the 39th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Florence, Italy, 2014, pp. 5502–5506.

[212] E. Yılmaz, J. F. Gemmeke, D. Van Compernolle, and H. Van hamme. “Noise-robust Digit Recognition with Exemplar-based Sparse Representations of Variable Length”. In: *IEEE Workshop on Machine Learning for Signal Processing (MLSP)*. Santander, Spain, 2012.

[213] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. *The HTK Book Version 3.3*. Cambridge University Press, 2005.

[214] D. Yu, L. Deng, and A. Acero. “A lattice search technique for a long-contextual-span hidden trajectory model of speech”. In: *Speech Communication* 48.9 (2006), pp. 1214–1226.

[215] H.-B. Yu and M.-W. Mak. “Comparison of Voice Activity Detectors for Interview Speech in NIST Speaker Recognition Evaluation”. In: *Proceedings of the 12th Annual Conference of International Speech Communication Association (INTERSPEECH)*. Florence, Italy, 2011, pp. 2353–2356.

[216] R. Zdunek and A. Cichocki. “Nonnegative matrix factorization with constrained second-order optimization”. In: *Signal Processing* 87.8 (2007), pp. 1904–1916.

[217] L. Zhang. “Modelling Speech Dynamics with Trajectory-HMMs”. PhD thesis. University of Edinburgh, 2009.

[218] S. Y. Zhao, S. Ravuri, and N. Morgan. “Multi-Stream to Many-Stream: Using Spectro-Temporal Features for ASR”. In: *Proceedings of the 34th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Taipei, Taiwan, 2009, pp. 2951–2954.

Corrections to Publications

The SDR measurements in [P4] were based on calculating difference signals between clean, noisy, and estimated utterances as in [45]. However, it was later found out that different magnitude scalings appear in the corpus depending on the data set. Either 'embedded' or 'isolated' utterances were used for experiments, causing inaccuracy within and across methods. For more representative measurement, the results have been recalculated here using the BSSeval toolkit [182].

The directly computed results as published in Table 4 of [P4] were as follows.

SNR	9 dB	6 dB	3 dB	0 dB	-3 dB	-6 dB	avg
Unenhanced signals and informed noise modelling							
Unenhanced	3.7	2.5	0.3	-1.9	-4.8	-7.0	-1.2
Informed	4.4	4.1	3.8	3.5	3.1	2.7	3.6
Self-adapting noise, all/true/estimated identity							
All	8.6	7.8	6.8	5.9	4.7	3.9	6.3
True	6.9	6.4	6.0	5.5	4.9	4.4	5.7
Estimated	6.9	6.4	6.0	5.4	4.6	4.0	5.6

Results for the same experiments, recomputed with BSSeval, are listed below.

SNR	9 dB	6 dB	3 dB	0 dB	-3 dB	-6 dB	avg
Unenhanced signals and informed noise modelling							
Unenhanced	4.3	3.0	0.6	-1.5	-4.4	-6.5	-0.8
Informed	13.1	11.3	9.5	7.8	5.9	4.3	8.6
Self-adapting noise, all/true/estimated identity							
All	10.7	9.4	7.8	6.3	4.5	3.5	7.0
True	9.9	9.0	8.2	7.2	6.1	5.1	7.6
Estimated	9.9	9.0	8.2	6.9	5.4	4.0	7.2

The observations and conclusions given in Sections 4.5 and 5 of [P4] no longer reflect these newly calculated values in every respect.

Publication P1

A. Hurmalainen, J. F. Gemmeke, and T. Virtanen, “Non-negative Matrix Deconvolution in Noise Robust Speech Recognition”, in *Proceedings of the 36th IEEE International Conference on Audio, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, 22.–27. May 2011, pp. 4588–4591.

NON-NEGATIVE MATRIX DECONVOLUTION IN NOISE ROBUST SPEECH RECOGNITION

Antti Hurmalainen* *Jort Gemmeke*† *Tuomas Virtanen**

* Tampere University of Technology, P.O. Box 553, 33101 Tampere, Finland
† Radboud University, Nijmegen, P.O. Box 9102, 6500 HC Nijmegen, The Netherlands

ABSTRACT

High noise robustness has been achieved in speech recognition by using sparse exemplar-based methods with spectrogram windows spanning up to 300 ms. A downside is that a large exemplar dictionary is required to cover sufficiently many spectral patterns and their temporal alignments within windows. We propose a recognition system based on a shift-invariant convolutive model, where exemplar activations at all the possible temporal positions jointly reconstruct an utterance. Recognition rates are evaluated using the AURORA-2 database, containing spoken digits with noise ranging from clean speech to -5 dB SNR. We obtain results superior to those, where the activations were found independently for each overlapping window.

Index Terms— Automatic speech recognition, noise robustness, deconvolution, sparsity, exemplar-based

1. INTRODUCTION

Widespread adoption of Automatic Speech Recognition (ASR) systems is still being hampered by insufficient robustness against background noise. Hidden Markov Model (HMM) based recognisers, where state likelihoods are estimated using Gaussian Mixture Models (GMM), have considerable problems when noisy frames no longer match to clean acoustic models. Various robustness methods have been suggested, including model compensation, missing data techniques and feature enhancement [1, 2, 3]. These approaches can typically achieve acceptable recognition rates in low to medium noise, but lose quality rapidly, when a large portion of spectral features is simultaneously corrupted by high noise levels.

In our previous work we have shown, that improved recognition rates can be achieved near or below 0 dB SNR by using an additive model of *exemplars* representing longer (100 – 300 ms) spectrogram segments [4]. Using a Non-negative Matrix Factorisation (NMF) algorithm, it is possible to separate the input signal to speech and noise. Furthermore, we have shown that speech content can be decoded directly from the labels of activated exemplars without reconstructing the separated speech signal [5].

In contrast to earlier exemplar-based methods, where the observation is compared to the nearest element in the dictionary, our framework reconstructs observations as a non-negative linear combination of exemplars. The number of simultaneously active exemplars is not limited by the design, although sparsity is enforced to improve the recognition quality. Similar methods have been used for source separation in image and music applications, among others. Common terminology for referring to such techniques includes *Sparse Classification* (SC) and *Sparse Representation based Classification* (SRC).

While the noise robustness of our algorithm improved by using longer exemplars, we also observed a decrease in clean speech recognition rates. The primary reason for this negative development is that the complexity of spectro-temporal features will increase in longer windows, thus requiring more exemplars to cover the larger variation in appearing patterns [6]. In addition, factorisation of individual analysis windows requires that correctly time-aligned exemplars are available in the dictionary, so the number of different temporal alignments to be covered also increases according to window length. However, simultaneous increasing of both exemplar count and length is not desirable due to computational constraints.

To improve the recognition accuracy of our system using a limited dictionary of long exemplars, we introduce a shift-invariant convolutive model. By reconstructing the whole observation at once as a convolution of exemplars and activations, we avoid the problem of temporal alignment of the exemplars in fixed windows. It is no longer necessary to include multiple shifted variants of features in the exemplars to represent the observation accurately. Consequently, better efficiency can be expected for similar dictionary size.

The content is organised as follows. Section 2 describes the key concepts of the paper: exemplar-based recognition, matrix deconvolution and differences to the previous model. In Section 3 we explain, how to obtain state likelihoods and the final recognition output from exemplar activations. The noisy spoken digit recognition test setup is given in Section 4. Results, discussion, and conclusions follow in Sections 5, 6 and 7, respectively.

2. EXEMPLAR-BASED DECONVOLUTION

2.1. Windowed exemplar model

The basis unit of our system, a speech or noise *exemplar*, is a $B \times T$ spectrogram matrix consisting of spectral magnitudes (square root of energy). B is the number of frequency bands and T the number of consecutive frames in each exemplar. Our observation matrix $\mathbf{Y}_{\text{utt}}$ is a $B \times T_{\text{utt}}$ spectrogram in the same domain, where T_{utt} is the total number of frames in the whole speech utterance.

The utterance is modelled as a linear weighted combination of exemplars in overlapping, exemplar-sized *windows*. The starting frame indices τ of windows range from 1 to $W = T_{\text{utt}} - T + 1$, and a window starting from frame τ covers frames $[\tau, \tau + T - 1]$. The linear combination is characterised by an $L \times W$ activation matrix $\mathbf{X}$, where each element X_{lw} represents the weight of exemplar l (from 1 to the total number L) activation in window w. The activation pattern can be determined for one window at a time as in our previous experiments [4, 5], or by generating joint activations for the whole utterance using a *deconvolution* algorithm.

Tuomas Virtanen and Antti Hurmalainen have been funded by the Academy of Finland. The research of Jort Gemmeke was carried out in the MIDAS project, granted under the Dutch-Flemish STEVIN program.

2.2. Matrix deconvolution

The estimated model $\mathbf{\Psi}_{\text{utt}}$ for observation $\mathbf{Y}_{\text{utt}}$ using L exemplars can be written as

$$\mathbf{\Psi}_{\text{utt}} = \sum_{t=1}^{T} \mathbf{A}_t \overset{\rightarrow(t-1)}{\mathbf{X}}. \qquad (1)$$

Each $\mathbf{A}_t$ is a $B \times L$ matrix representing frame t of the exemplars, thus the spectrogram of exemplar l can be found in columns l of $\mathbf{A}_1 \dots \mathbf{A}_T$. Here $\overset{\leftarrow i}{(\cdot)}$ and $\overset{\rightarrow i}{(\cdot)}$ are shift operators, moving the matrix entries left or right, respectively, by i units. In this case $\mathbf{\Psi}_{\text{utt}}$ is $T-1$ columns longer than the activation matrix $\mathbf{X}$, so shifting takes place in a T_{utt} wide zero-padded matrix, starting from its leftmost position. $T-1$ zero columns are added, no columns are discarded to generate the shifted matrix.

The exemplars and their activations are restricted to non-negative values. The exemplars are obtained from training data and fixed, whereafter the activations are estimated by minimising the generalised Kullback-Leibler divergence

$$d(\mathbf{Y}_{\text{utt}}, \mathbf{\Psi}_{\text{utt}}) = \sum y \log(\frac{y}{\psi}) - y + \psi \quad \forall (y, \psi) \in (\mathbf{Y}_{\text{utt}}, \mathbf{\Psi}_{\text{utt}}). \qquad (2)$$

An L_1 norm penalty (sum of all elements) is applied to the activations, which has been found effective for magnitude spectrogram features [7].

As the approximated observation matrix $\mathbf{\Psi}_{\text{utt}}$ will be a temporal convolution between the basis and the activations, the algorithm is called Non-negative Matrix Deconvolution (NMD) [8]. In our previous work we called the method *convolutive sparse coding* [9]. NMD has already been used successfully for sound source separation in music and speech applications [10, 11].

The entries of the activation matrix are initialised to unity values, and the following update rule (based on [12]) is applied iteratively:

$$\mathbf{X} = \mathbf{X} \otimes \frac{\sum_{t=1}^{T} \mathbf{A}_t^T \cdot \overset{\leftarrow(t-1)}{[\frac{\mathbf{Y}_{\text{utt}}}{\mathbf{\Psi}_{\text{utt}}}]}}{\mathbf{\Lambda} + \sum_{t=1}^{T} \mathbf{A}_t^T \cdot \overset{\leftarrow(t-1)}{\mathbf{1}}}, \qquad (3)$$

where $\otimes$ is elementwise multiplication, and all divisions are also elementwise. $\mathbf{\Lambda}$ is a sparsity matrix defining the penalty factor for each activation element, thus the total weighted penalty becomes $\sum x \cdot \lambda \quad \forall (x, \lambda) \in (\mathbf{X}, \mathbf{\Lambda})$. In our system, we set a different penalty weight for activations corresponding to speech and noise. The model $\mathbf{\Psi}_{\text{utt}}$ is evaluated before each update using (1).

2.3. Comparison to independent windows

In our previous work we used a sliding window approach, where all W overlapping $B \times T$ windows were factorised independently. Because the middle frames of the observation will be reconstructed several times in consecutive windows, averaging was applied in later steps to compensate for the effect. The implementation was somewhat simpler than in NMD — each window can be represented as a separate, concatenated observation vector, and the utterance can be processed as a factorisation between two matrices without shifting operations. However, it occasionally suffers from the fixed temporal positioning of its windows. An exemplar must match accurately to the temporal position of spectral features found in an individual window to be used there. When the window length is increased, it becomes less likely, that a matching exemplar is found in a limited dictionary. Each window must be factorised, and depending on its match to the dictionary, reconstruction quality may vary between windows. The effect of mismatches will be reduced during averaging, but not eliminated entirely. On the other hand, for NMD it suffices to find a single temporal position, where an exemplar matches the observed speech. The difference between the activation patterns is visualised in Figure 1.

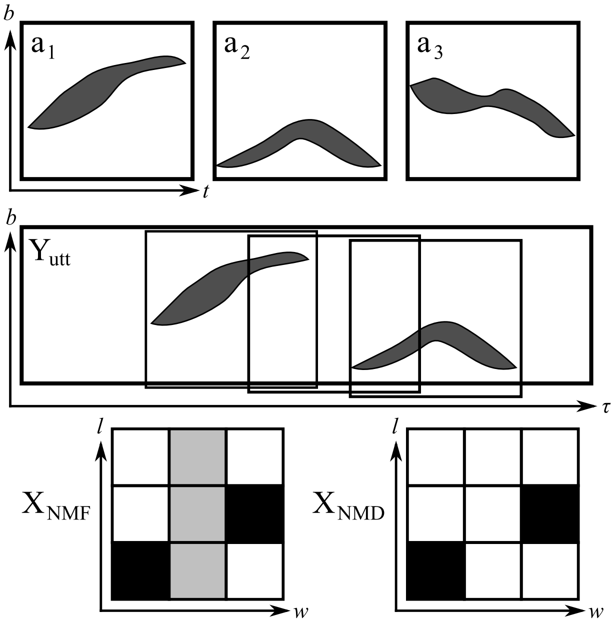


Figure 1: A stylised comparison of independent window (NMF for short) and deconvolution (NMD) methods. Utterance spectrogram $\mathbf{Y}_{\text{utt}}$ is represented using exemplars $\mathbf{a}_1$, $\mathbf{a}_2$ and $\mathbf{a}_3$ in three windows. The first and last window match to exemplars 1 and 2, but in NMF the middle window must be reconstructed using inaccurate activations (bottom left matrix). In NMD, only enough exemplars to reconstruct the utterance are activated (bottom right matrix), thus the middle window remains empty.

3. DECODING

After determining the activation matrix $\mathbf{X}$, it is used to generate a state likelihood matrix $\mathbf{L}$. It consists of column vectors $\mathbf{l}_\tau$ for each frame in the utterance. These vectors, their length representing the total number of states in the system, describe the estimated likelihoods of states at time τ.

Each speech exemplar is labelled with a state sequence over its duration, so that in each frame it is assumed to be in exactly one state. When an exemplar is activated in window w, an update is made to T columns of $\mathbf{L}$ starting from w. A state label q in frame t of an exemplar will increment the element q of column $w+t-1$ by its activation weight. A formal description of this procedure is given in [5].

Even though silence states are also included in the labels, their activation is somewhat unpredictable. Because the magnitude of silent frames is zero in all bands, no exemplars are activated during true silence. Conversely, these states may appear within speech activity, when a speech-silence transition exemplar is used as a part of the sum. For these reasons, silence state likelihoods are reshaped according to a speech activity estimate derived from the total weight of active speech exemplars in each frame. The matter is discussed in

Table 1: Digit recognition rates for AURORA-2 test sets A and B at various window lengths and noise levels. The first three rows repeat the independent window factorisation ('NMF' for short) results given in [5]. The last three rows show the new deconvolution results ('NMD').

SNR	(dB)	clean	20	15	10	5	0	-5
NMF	T=10	96.2	95.3	94.4	92.1	84.7	71.2	39.6
	T=20	96.6	95.8	94.8	92.7	88.8	78.1	53.1
	T=30	94.7	93.4	93.3	92.2	89.9	79.5	56.7
NMD	T=20	96.7	96.3	95.4	93.9	90.1	78.5	57.5
	T=30	97.0	96.4	95.6	94.7	91.4	82.0	61.0
	T=40	93.5	94.4	94.2	91.5	88.6	78.3	55.2

(a) Test set A

SNR	(dB)	clean	20	15	10	5	0	-5
NMF	T=10	96.2	94.7	93.6	87.9	78.4	57.1	27.4
	T=20	96.6	95.3	93.7	89.9	82.7	63.1	35.7
	T=30	94.7	93.5	93.2	90.1	85.7	67.5	37.6
NMD	T=20	96.7	96.0	95.1	91.7	84.0	62.4	33.5
	T=30	97.0	95.6	94.7	92.1	86.4	68.1	36.4
	T=40	93.5	93.8	93.4	89.2	83.6	64.1	33.0

(b) Test set B

more detail in [4].

Finally, the summed likelihoods in each frame are normalised to unity, and the state likelihood matrix is decoded using the Viterbi algorithm.

4. EXPERIMENTS

The efficiency of deconvolution versus independent overlapping windows was studied using a test setup similar to the one described in [4] and [5]. AURORA-2 connected digit recognition test, which includes multiple noise types and noise levels, was used for evaluation. The same bases of 4000 speech and 4000 noise exemplars, generated by random selection from the multicondition training set in the earlier experiments, were used. In these bases, each exemplar is a $B \times T$ magnitude spectrogram consisting of 23 mel-scale spectral bands and T frames with 25 ms frame length and 10 ms frame shift. Window lengths 20 and 30 from the previous work were included, as this much temporal context has been found recommendable for sufficient noise robustness. In addition, a $T = 40$ basis was generated using a similar procedure to study the capability of deconvolution in even longer windows. State labels of speech exemplars were acquired via HMM-based forced alignment. All in all, 179 states were used: 16 for each digit ('zero', 'oh', 1–9) and 3 for silence.

We processed the same random subset of 100 utterances (10% of the complete test set) as in [5] for all four noise types in test set A and the four in test set B. Clean speech and all six noise levels, SNR 20, 15, 10, 5, 0 and -5 dB were included. Due to the different activation patterns between independent windows and deconvolution, the NMD sparsity parameters λ were reoptimised to 2.0 for speech and 1.5 for noise exemplars using the training set. The silence balancing algorithm was modified slightly to derive its SNR estimate from waveforms by comparing the mean power of the whole wave (signal+noise) to the lowest 20% of frame powers (only noise), because in NMD the exemplar activation levels were found to vary too much for this purpose. The silence parameters were retrained from the training set for each window length separately. 200 NMD iterations were used for the main experiment as before, although the computation was continued up to 250 iterations for further comparison.

5. RESULTS

The recognition rates of our test are summarised in Table 1. Previous results from our independent window experiments ('NMF') are shown first, sorted by window length [5]. The new convolutive model ('NMD') results follow.

In set A, convolutive $T = 30$ comes out uniformly superior to the alternative window lengths and also to our previous results. Convolutive $T = 20$ surpasses the NMF results and approximately equals convolutive $T = 30$ at high SNRs, but falls faster in the noisy end like it did in NMF. The newly introduced $T = 40$ (400 ms exemplars) is roughly comparable to the previous $T = 20/30$ NMF results. However, a decrease of approximately 3% from convolutive $T = 30$ is present already in the clean end, and it reflects to all the noisy rates. Overall, set A turns out to be a success for the convolutive algorithm.

In set B we observe mostly positive results, but also a few decrements. The improvements in clean speech recognition rates are also present here all the way until 0 dB, where convolutive $T = 20$ loses by a small margin to its NMF counterpart. For $T = 30$, this happens at -5 dB alone. $T = 40$ is again acceptable in comparison to the NMF results, but several percent below the new $T = 30$ rates.

The high contrast between set A (noise types matching to the basis) and set B (nonmatching noise) is still present and even emphasised in the convolutive approach. The possible reasons for this are discussed in Section 6.

Increasing the iteration count to 250 produced mixed results (not shown). Recognition rate changes between -1.4% and +3.7% (absolute) were observed. The largest and most systematic gains were in the noisy end of set A, all 0 dB rates increasing by $\geq 1.0\%$ and -5 dB by $\geq 2.2\%$. Elsewhere no regular trend was found.

In comparison to established methods, the current experimental setup does not yet achieve the clean speech recognition rates of carefully trained GMM-based implementations, which often exceed 99%. On the other hand, previous -5 dB rates achieved with noise-compensated or multi-condition trained GMMs include 17.1% [2], 24.6% [13] and 42.9% [4] for set A. All perform worse on set B, albeit by a smaller margin, when the methods do not utilise spectro-temporal features specific for each noise type. Uncompensated systems trained with clean speech typically fall below 10% at -5 dB.

6. DISCUSSION

Three main observations can be made from the results. First, in this test setup the convolutive method produces generally higher recognition rates than the independent window algorithm. Second, convolutive $T = 30$ achieves the highest clean speech recognition rate of all methods and windows presented here, improving significantly its earlier independent window performance. Third, test set B still turns out problematic, even more so than in NMF. Each of these observations deserves a brief analysis.

The improved overall rates are a positive outcome, and speak for the potential of NMD in exemplar-based recognition. However, the new algorithm also required some changes and retraining of parameters, which may play a role in the overall results. We still conclude, that significant gains were achieved by using NMD for the problem.

Because its joint, shift-invariant activation pattern appears inherently suitable for dictionary reduction and reverberation handling, we consider it the better candidate for further research within related topics, such as echoing noise and large vocabulary.

The second observation was the superiority of $T = 30$. Whereas in the previous independent window experiment it suffered from lower clean speech recognition rates, here it improves to the extent that it surpasses both of the $T = 20$ variants in all SNRs. It was our earlier assumption, that in such a long window the dictionary size becomes a limiting factor for independent windows, because several temporal alignments of features are required in the exemplars. We also assumed, that deconvolution might reduce the effect. The results support both of these theories. As the $T = 30$ basis was identical in both variants, and post-processing factors are negligible in clean speech recognition, we conclude that the nearly halved error rate in clean speech results from algorithmic differences. The other high percentages in set A follow the improved performance of clean speech throughout the noise levels. Window length 40 was found too large to be handled with this dictionary size, regardless of the use of convolution.

The primary problem of our current approach is highlighted by the third observation, namely the increasing quality gap between sets A and B. The noise types of set A are similar to those used in training and dictionary construction. Therefore the factorisation/deconvolution becomes a well defined separation problem, and generally plausible results can be achieved. The situation is notably different in test set B. Because the noise types do not match, especially in long windows we cannot expect to find good approximations for the observed noise in the dictionary. In NMF of independent windows, a lot of averaging will take place. Up to 30 different noise estimates from consecutive windows are mixed together. Therefore they are unlikely to form any major distracting features. In NMD, this kind of forced averaging is not present. The increased sparsity, which aided separation in set A, may become a hindrance instead. Sparse activations of nonmatching noise features are not suitable for representing the true noise in signals, thus the separation often fails. A telling detail is that in set A the noisy results improved further by increasing the iterations to 250. In set B this did not happen. Even a few decrements took place, suggesting that the algorithm had already reached an unstable peak level regarding separation quality.

It has been repeatedly seen that long temporal context is effective, or even required for handling high levels of background noise. We also found here additional support for the potential of exemplar-based sparse representation. However, while various speech patterns can be handled by a reasonably sized exemplar dictionary, the same cannot be said about all types of noise present in the real world. To cope with this issue, we have already taken initial steps towards adaptive and synthetic noise dictionaries [14]. Preliminary results show that even a simple synthetic dictionary can surpass the separation quality of a poorly matching sampled dictionary. Deconvolution should prove useful in such dictionary methods, because new patterns can be included as single entries without temporal repetition. The algorithm itself will take care of different temporal alignments.

7. CONCLUSIONS

A framework for an exemplar-based, deconvolutive speech recognition system was presented. Comparative results against an earlier setup with independent factorisation windows were shown using the AURORA-2 connected digit recognition test. Deconvolution with a window length of 30 frames (300 ms) surpassed the results of other window lengths and the previous approach almost uniformly. Recognition rates of >80% were observed at 0 dB SNR, and >60% at -5 dB. Improvements in clean speech recognition rates using long windows suggest, that deconvolution can overcome some of the dictionary size problems of independent windows. It turned out that the match between the dictionary and observed noise is crucial in deconvolution, even more so than in the independent window approach.

8. REFERENCES

[1] A. Acero, L. Deng, T. Kristjansson, and J. Zhang, "HMM adaptation using vector Taylor series for noisy speech recognition," in *Proceedings of ICSLP*, 2000, pp. 869–872.

[2] H. Van hamme, "Robust speech recognition using cepstral domain missing data techniques and noisy masks," in *Proceedings of ICASSP*, 2004, pp. 213–216.

[3] B. Raj and R.M. Stern, "Missing-feature approaches in speech recognition," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 101–116, September 2005.

[4] J.F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *Accepted for publication in IEEE Transactions on Audio, Speech, and Language Processing*, 2011.

[5] T. Virtanen, J.F. Gemmeke, and A. Hurmalainen, "State-based labelling for a sparse representation of speech and its application to robust speech recognition," in *Proceedings of INTERSPEECH*, 2010.

[6] J. Gemmeke, L. ten Bosch, L. Boves, and B. Cranen, "Using sparse representations for exemplar based continuous digit recognition," in *Proceedings of EUSIPCO*, 2009, pp. 1755–1759.

[7] T. Virtanen, "Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, 2007.

[8] P. Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs," in *Independent Component Analysis and Blind Signal Separation*, pp. 494–499. 2004.

[9] T. Virtanen, "Separation of sound sources by convolutive sparse coding," in *Proceedings of ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing*, 2004.

[10] P.D. O'Grady and B.A. Pearlmutter, "Discovering speech phones using convolutive non-negative matrix factorisation with a sparseness constraint," *Neurocomputing*, pp. 88–101, 2008.

[11] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, 2007.

[12] T. Virtanen, *Sound source separation in monaural music signals*, Ph.D. thesis, Tampere University of Technology, 2006.

[13] H. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proceedings of ISCA Tutorial and Research Workshop*, 2000, pp. 181–188.

[14] J.F. Gemmeke and T. Virtanen, "Artificial and online acquired noise dictionaries for noise robust ASR," in *Proceedings of INTERSPEECH*, 2010.

Publication P2

A. Hurmalainen, K. Mahkonen, J. F. Gemmeke, and T. Virtanen, “Exemplar-based Recognition of Speech in Highly Variable Noise”, in *Proceedings of the 1st International Workshop on Machine Listening in Multisource Environments (CHiME)*, Florence, Italy, 1. September 2011, pp. 1–5.

Exemplar-based Recognition of Speech in Highly Variable Noise

Antti Hurmalainen[1], *Katariina Mahkonen*[1], *Jort F. Gemmeke*[2], *Tuomas Virtanen*[1]

[1]Department of Signal Processing, Tampere University of Technology, Finland
[2]Department ESAT, Katholieke Universiteit Leuven, Belgium
antti.hurmalainen@tut.fi, katariina.mahkonen@tut.fi,
jgemmeke@amadana.nl, tuomas.virtanen@tut.fi

Abstract

Robustness against varying background noise is a crucial requirement for the use of automatic speech recognition in everyday situations. In previous work, we proposed an exemplar-based recognition system for tackling the issue at low SNRs. In this work, we compare several exemplar-based factorisation and decoding algorithms in pursuit of higher noise robustness. The algorithms are evaluated using the PASCAL CHiME challenge corpus, which contains multiple speakers and authentic living room noise at six SNRs ranging from 9 to -6 dB. The results show that the proposed exemplar-based techniques offer a substantial improvement in the noise robustness of speech recognition.

Index Terms: automatic speech recognition, exemplar-based, noise robustness, sparse representation

1. Introduction

While Automatic Speech Recognition (ASR) has been under intensive research for decades, its widespread adoption is still being delayed by practical issues. One of the primary problems is varying background noise. Conventional ASR systems, based on frame level Gaussian Mixture Models (GMMs), suffer significant quality degradation when spectral features become corrupted by noise. Joint modelling of the target speech and noise in the recognizer, [1], feature compensation [2], and missing data techniques [3] have been suggested to overcome this problem. Meanwhile, there are alternative routes, which no longer employ GMMs to discover the underlying speech content.

In previous work [4, 5], we have described an *exemplar-based* recognition framework, where noisy speech is represented as a combination of multi-frame speech and noise spectrogram segments, *exemplars*. The framework can be used for signal or feature enhancement, but the best results have been achieved by using exemplar labels, which directly reveal the phonetic content of an utterance via their activation weights. In this paper, we explore the effectiveness of the exemplar-based framework on highly corrupted speech using the PASCAL CHiME challenge data, in which the speech is not only reverberated, but also contains phonetically close keywords and highly variable background noise events.

Concerning our framework, the CHiME data provides a few interesting options, which were not present in the previous experiments carried out on the AURORA-2 database. First, the data is stereophonic and high quality. Second, the utterances to be recognised can be observed within their neighbouring noise context. Finally, the identity of the speaker is known at the moment of recognition, so speaker-dependent speech exemplars can be reliably employed.

The rest of the paper is organised as follows. The general concepts of our exemplar-based approach are described in Section 2. The experimental setup, including the CHiME database, feature extraction and parameter settings of the baseline system are presented in Section 3. The baseline exemplar-based recognition results are shown and discussed in Section 4. Experiments with two variants; the use of matrix deconvolution (NMD) and the use of regression to learn the mapping between words and exemplars, are described in Sections 5 and 6, respectively. The overall discussion of our findings is presented in Section 7, and the summary and conclusions in Section 8.

2. Recognition with speech and noise exemplars

Sparse representations have received increasing attention in several applications, including image and audio signal processing. The key concept is that many natural signals can be described as a linear combination of only a few atoms. Enforcing sparsity prevents overfitting with too many elements. By allowing only a small number of activations, we can expect to find the few dictionary atoms, which best explain the mixed signal.

In noise robust speech recognition, it has been proposed that speech may be described as a sparse linear combination of *exemplars*, and that noisy speech can likewise be described as a combination of noise and speech exemplars [5, 6, 7]. When a noisy utterance is represented using these components, the activations of speech exemplars, together with knowledge of the words they represent, can be used to recognise the underlying utterance.

2.1. Sparse representation of noisy speech

The base element of our sparse representation is an *exemplar*, a $B \times T$ spectrogram block of B spectral magnitudes of speech or noise in T consecutive frames, extracted from training data. The exemplars are indexed by variable e. To simplify the notation, the columns of each spectrogram matrix are stacked into vector $\mathbf{a}_e$ of length $B \cdot T$. The E exemplars are gathered into the columns of matrix $\mathbf{A}$ to form a *basis* or *dictionary*.

The utterance to be recognised is similarly converted to spectral features. A length T observation window is concatenated into vector $\mathbf{y}$. The observation window is represented as a linear combination of exemplars,

$$\mathbf{y} \approx \sum_{e=1}^{E} \mathbf{a}_e x_e, \tag{1}$$

where x_e is the weight or *activation* of each exemplar.

In the baseline exemplar-based recognition system we employ an algorithm referred as 'NMF' (*Non-negative Matrix Factorisation*) to find the non-negative and sparse activations. The vector $\mathbf{x}$ of all activations x_e in Equation 1 can be determined simultanously for multiple observation vectors stored in columns of matrix $\mathbf{Y}$, each producing its own column to the total activation matrix $\mathbf{X}$. The matrix equation to be solved thus becomes $\mathbf{Y} \approx \mathbf{AX}$.

We obtain the non-negative activation matrix $\mathbf{X}$ while minimising the Kullback-Leibler divergence and introducing an sparsity-inducing L_1 penalty for non-zero activations by using the update rule

$$\mathbf{X} \leftarrow \mathbf{X} \otimes \frac{\mathbf{A}^{\mathrm{T}}(\mathbf{Y}/(\mathbf{AX}))}{\mathbf{A}^{\mathrm{T}}\mathbf{1} + \mathbf{\Lambda}}. \tag{2}$$

Here $\otimes$ denotes elementwise multiplication. Matrix divisions are also elementwise. $\mathbf{1}$ is an utterance-sized all-ones matrix. $\mathbf{\Lambda}$ is the sparsity penalty matrix, defined for each activation entry.

For recognition of utterances of arbitrary length T_{utt}, we process the utterance in $W = T_{\text{utt}} - T + 1$ overlapping feature windows with a step of one frame between windows. Because the middle frames are estimated several times in consecutive windows, averaging is applied to the likelihoods of the next step to compensate for this. For a thorough description of this factorisation method, see [4]. An alternative method for handling temporal continuity, referred as *Non-negative Matrix Deconvolution* (NMD), is presented in Section 5.

2.2. Recognition

To decode the signal, we create a $Q \times T_{\text{utt}}$ *likelihood matrix* $\mathbf{L}$, where each entry $\mathbf{L}_{q\tau}$ denotes the probability of speech state q $(1 \ldots Q)$ in frame τ $(1 \ldots T_{\text{utt}})$. This is generated using *conversion matrices* $\mathbf{B}_t$ $(Q \times E)$, which describe the linear mapping of exemplars to states for each frame t of the exemplars. In our baseline system, we use binary labelling of dictionary exemplars for the conversion. In each exemplar frame only one state is labelled to be active. The matrices need not to be binary, though. in Section 6 we will experiment with a technique to *learn* the conversion matrices in order to take into account dependencies between exemplar activations.

After generating the whole matrix $\mathbf{L}$ as described in [4], each of its columns (representing state likelihoods in one frame) is normalised to unitary sum. The matrix is then decoded using a Viterbi algorithm and trained transition parameters.

3. Experimental setup

3.1. The CHiME database

The PASCAL 'CHiME' Speech Separation and Recognition Challenge [8] is designed to address some of the problems occurring in real world noisy speech recognition. The challenge data is based on the GRID corpus [9], where 34 speakers read simple command sentences. These sentences are of form *verb-colour-preposition-letter-digit-adverb*. There are 25 different 'letter' class words and 10 different digits. Other classes have four word options each. In the CHiME recognition task, the final score is the percentage of correctly recognised 'letter' and 'digit' keywords.

CHiME utterances simulate a scenario, where sentences are spoken in a noisy living room. The original, clean speech utterances are reverberated according to the actual room response, and then mixed to selected noise sections, which produce the desired SNR mixture level for each noisy set. The noisy sets have target SNR levels of 9, 6, 3, 0, -3 and -6 dB.

For modelling/training, there are 500 reverberated utterances per speaker (no noise), and six hours of background noise data. The development and test sets consist of 600 mixed-speaker utterances at each SNR level, Additionally, noiseless (only reverberated) development utterances are available. Development and test utterances are both given in a strictly end-pointed format, but also as embedded signals within their noise context. All data is stereophonic and has a sampling rate of 16 kHz.

3.2. Feature extraction

For the features of our framework, we used spectral magnitudes of Mel bands. These were calculated from partially overlapping 25 ms frames with a shift of 10 ms between frames. 26 bands were used for the 16 kHz signal (Nyquist frequency 8 kHz), which matches the number of bands used for the default CHiME MFCC models. Features were extracted separately for both stereo channels and concatenated, thus effectively doubling the number of feature bands.

3.3. Speech exemplars

We used 5000 speech and 5000 noise exemplars for each window length T, adding up to $E = 10000$ total entries. We created two different types of speech dictionaries: a speaker-dependent and a speaker-independent one. First, an initial speech dictionary was created for each speaker, based on a 60% subset of the noiseless speech training utterances, by extracting exemplars with a random frame shift of 4 to 8 frames. This produced approximately 10000–17000 partially overlapping exemplars per speaker and window length. For the speaker-dependent dictionaries, each initial dictionary was reduced to a fixed size of 5000 exemplars by selecting exemplars such that there is a maximally flat coverage between words. (In the original dictionaries, words from classes with fewer options are over-represented due to more frequent appearance in the training set.)

A speaker-independent dictionary was created for each window length, this time by selecting 147–148 (5000/34) exemplars from each full speaker-dependent dictionary with similar word probability flattening. These were then combined to a single 5000 exemplar dictionary per window length.

In addition to storing the spectral feature data, state labels were assigned to the speech exemplars by using transcriptions acquired by forced alignment. Alternatively, the state information was learnt by factorising the remaining 40% of training files and finding the mapping as described in Section 6.

3.4. Noise exemplars

The selection of noise exemplars has a central role in the separation quality of factorisation algorithms. If no matching noise is found, separation results become unpredictable. Initially, we created two different types of noise dictionaries. In the first, 5000 noise exemplars were randomly extracted from the provided background noise data. In the second, 5000 noise exemplars were selected by sampling the neighbourhood of embedded utterances to both directions with a shift of 4 to 7 frames, excluding locations where other test utterances were embedded.

Experiments using the development set (not shown) indicated that using the adaptive noise dictionary yields a 1–4% improvement in recognition accuracy compared to the fixed noise dictionary. In this paper, we will only report results obtained using adaptive noise.

Table 1: Results of the baseline exemplar-based recogniser on the test set. The rows refer to different exemplar sizes. CHiME GMM baseline results are also shown. The best result at each SNR level is highlighted.

SNR (dB)	9	6	3	0	-3	-6
CHiME baseline	**82.1**	70.8	61.3	52.0	39.8	34.7
$T = 10$	69.9	66.0	58.7	52.4	42.9	37.8
$T = 20$	77.3	72.8	**68.2**	**62.7**	51.1	44.0
$T = 30$	76.0	**73.5**	**68.2**	61.8	**52.7**	**44.7**

(a) Speaker-independent results

SNR (dB)	9	6	3	0	-3	-6
CHiME baseline	82.4	75.0	62.9	49.5	35.4	30.3
$T = 10$	91.3	88.3	85.8	80.8	71.4	62.3
$T = 20$	**91.6**	**89.2**	**87.6**	**84.2**	74.7	68.0
$T = 30$	88.8	88.1	86.3	82.9	**75.1**	**68.3**

(b) Speaker-dependent results

3.5. Processing test utterances

For factorisation, each utterance was read from the endpointed ('isolated') file, and converted into Mel features. After choosing the appropriate speech and noise basis for the utterance, they were reweighted together to equal Euclidean norm over Mel bands and exemplars. Band weights from the combined dictionary were then applied to the utterance features.

The NMF penalty matrix $\mathbf{\Lambda}$ used in finding a sparse representation can be set for each activation entry separately. We used two different values, one for speech exemplars and another for noise. The values were tuned by factorising a subset of development utterances with partially adaptive, speaker-dependent bases and exemplar size $T = 20$. The penalty values were set as 2.0 and 1.7 for speech and noise exemplars, respectively. Generally speaking, higher values of $\mathbf{\Lambda}$ produce better recognition rates at high SNRs, while lower ones lead to better performance at low SNRs. We selected values, which give a slight emphasis to the noisy end. The same sparsity values were used throughout all experiments.

For state representation, we used the same model as in the CHiME baseline recogniser. Each word is modelled with 4–10 successive states, and the whole system uses in total 250 states. The activations were mapped to state likelihoods as explained in section 2.2. Utterances were decoded using the HVite binary of the HTK toolkit, modified to pick its state likelihoods directly from the generated matrix $\mathbf{L}$ instead of evaluating state GMMs.

4. Baseline system results

The results of the baseline exemplar-based recogniser are presented in Table 1. Three different window lengths, $T = 10$, 20 and 30 are shown, as well as results for both speaker-dependent and speaker-independent systems. The GMM-based CHiME baseline recognition results are also shown. When comparing the results, note that the baseline system uses mono features without noise compensation other than cepstral mean normalisation.

In general, it is clear that the exemplar-based recognition system outperforms the baseline GMM system in almost all conditions, especially when using speaker-dependent speech dictionaries. The lower performance of speaker-independent dictionaries ensues because a mixed speech dictionary only has a very limited number of exemplars to match a certain speaker, while at the same time it has a larger chance of matching to speech features in the background noise, produced by people in the living room or by various entertainment appliances. Interestingly, the speaker-independent GMM-based system was more noise robust at low SNRs, possibly because the trained Gaussians have a larger variance and thus match corrupted speech features better.

Like in experiments on AURORA-2 [4, 5], using an exemplar size of $T = 10$ was found suboptimal at low SNRs, because not enough time context can be exploited. $T = 20$ generally turned out equal or superior to $T = 10$. Exemplar size $T = 30$ is the most robust against noise, but performs worse at high SNRs. As the exemplar size increases, the dimensionality of feature vectors grows, and it becomes more difficult to find a matching linear combination of speech exemplars. Using a higher number of exemplars may alleviate this effect, at the cost of increased computational complexity.

5. Non-negative matrix deconvolution

As a first variant of the baseline exemplar-based recognition system, we use *Non-negative Matrix Deconvolution* (NMD) rather than NMF to obtain sparse representations of noisy speech. NMD is a name given to an alternative method to handle temporal continuity between frames. The algorithm has also been called *convolutive sparse coding* [10].

While not a deconvolution algorithm in the traditional sense, the name reflects the principle that a reconstructed utterance is represented as a convolution between activations and exemplars. This means that all the activations jointly form the estimated utterance matrix. A few activations at specific temporal locations are typically enough to represent the utterance features. There are no independent estimates or averaging like in the sliding window NMF. For the convolutive update algorithm and comparison of behaviour, see [11].

The results for NMF and NMD algorithms are shown in Table 2. Both methods employ adaptive noise dictionaries, speaker-dependent speech dictionaries and 300 iterative updates. In NMF, the speech exemplar activations were normalised to unitary sum in each window. In NMD, no normalisation was performed. These choices have been found recommendable in earlier work [4, 11].

In these results, NMF produces slightly yet significantly better recognition rates in all conditions. This is surprising, because on AURORA-2 we observed the opposite: NMD outperformed NMF. Especially the degradation of NMD at $T = 30$ is unexpected, because on AURORA-2 it was the best performing exemplar size [11].

One possible reason is that factorisation parameters were optimised using NMF. Because the full optimisation process is computationally heavy, the same parameters were applied directly to NMD. Therefore the results may favour NMF. We can also speculate, that the closely related keywords in CHiME are prone to occasional misclassifications in sparse activations. As there is more averaging over independent estimates in NMF, the chance of errors in the final estimate is smaller than in NMD. Because a 1–2% drop was already present in the cleanest end of both keyword classes, we can suspect a problem with word recognition itself, not the noise robustness of NMD.

Table 2: Comparison of NMF and NMD factorisation algorithms in speaker-dependent recognition. The rows refer to different exemplar sizes. The best result at each SNR level is highlighted.

SNR (dB)	9	6	3	0	-3	-6
CHiME baseline	82.4	75.0	62.9	49.5	35.4	30.3
$T = 10$	91.3	88.3	85.8	80.8	71.4	62.3
$T = 20$	**91.6**	**89.2**	**87.6**	**84.2**	74.7	68.0
$T = 30$	88.8	88.1	86.3	82.9	**75.1**	**68.3**

(a) NMF

SNR (dB)	9	6	3	0	-3	-6
CHiME baseline	82.4	75.0	62.9	49.5	35.4	30.3
$T = 10$	88.3	85.9	83.3	78.8	69.1	59.8
$T = 20$	**90.5**	**88.6**	**87.0**	**81.3**	**72.1**	**65.9**
$T = 30$	87.2	86.1	84.0	79.9	70.6	63.3

(b) NMD

6. Mapping from activations to likelihoods

In our baseline system, the mapping from activations to word state likelihoods is based on labels of dictionary items, which have been obtained by forced alignment. However, in label-based mapping of word models there is the inherent problem that phonetically similar features may appear in different contexts. A factorisation algorithm (NMF or NMD) selects the exemplars with best fitting spectral features, while their labels may occasionally suggest a misleading word identity. Such an error will easily result in a misclassification.

We tested an alternative approach, where the mapping was not assigned according to dictionary labels, but *learnt* using regression algorithms on factorised training data labelled by forced alignment. Labels were assigned to a 40% subset of the training set for this purpose. Then a regression algorithm was used to discover optimal mapping matrices between activation vectors and target states.

We used two different regression algorithms, *Ordinary Least Squares* (OLS) and *Partial Least Squares* (PLS) to learn the mapping from activations to likelihoods. OLS is straightforward minimisation of the L_2 error term in mapping. PLS (also known as Projection to Latent Structures) uses an internal, usually lower dimensioned space. The original coordinates are rotated in input and output to the internal space, where the true mapping is optimised. PLS can tolerate a collinearity of input data, contrary to OLS. For details, see [12].

The outcome of the recognition with different likelihood generation methods is shown in Table 3. Results are listed for recognition with binary labels, and OLS/PLS-trained mapping. Speaker-independent results are included, because they provide interesting insight to scenarios where flaws of the original system can be countered with learning.

In speaker-independent recognition, uniform improvements of 4.3–14.1% (absolute) can be seen over the use of binary labels. In these dictionaries, very few instances of each word are present for a specific speaker. This seems to result in numerous misclassifications due to exemplars from other words being activated instead. When the conversion matrices are learnt — in this case from a large amount of training material — the actual correspondence of each exemplar can be discovered with convincing results. Possibly for the abundance of training material coming from all speakers, OLS is mostly superior to PLS.

The speaker-dependent results are more mixed. Here the dictionaries only cover one speaker at a time, and thus can include a broad representation of all words and states. In fact, the reduction algorithm did not remove any of the letter and digit exemplars gathered from the training material, because they all fit in the 5000 exemplar dictionaries. It is also worth noting, that in this scenario the regression matrices were only trained from the speaker's own training subset (200 utterances), which is quite limited regarding keyword appearance. Under this limited training data, the performance of all methods was mostly similar, unlike in the speaker-independent case.

7. Discussion

The CHiME challenge database provided some new insight to the applicability of our exemplar-based methods. Overall, the results appear very plausible. Using properly selected algorithms and parameters, our framework reduced the recognition error rates to less than half of the CHiME baseline system at all SNRs, in many cases even by significantly larger a margin. We also achieved improvements in noise robustness over our previous work on AURORA-2. These gains can be partially attributed to the characteristics of CHiME, which allow construction of accurate dictionaries for both speech and noise.

When the speaker identity is known and thus matching speech exemplars can be selected, correct phonetic features can be picked out reliably even in the presence of other voices. Our speaker-dependent results were significantly better than the speaker-independent ones. Using GMMs the difference was not so clear. Regarding noise dictionaries, we found out that adaptive noise exemplar selection can yield high separation quality under varying noise conditions. Previously there were some concerns over the feasibility of generating a generic noise dictionary using a practically manageable number of exemplars. Our CHiME experiments confirm, that adaptive selection can be used instead of a fixed dictionary. Its implementation should be feasible in practical applications as well.

One surprising and slightly disappointing aspect was the subpar performance of NMD in comparison to sliding window based NMF. It is not certain yet, whether this is a real algorithmic difference or merely a result of insufficient parameter training in NMD. Further experiments and optimisations should be carried out to find out the true capabilities of each factorisation algorithm.

More favourable results were achieved in learnt likelihood mapping. The gains over explicitly assigned labels are positive by themselves. However, in a larger context this means that well performing likelihood mappings can be learnt even for features, which are not directly derived from any specific speech sections. In other words, we can experiment with any kind of dictionary generation methods and then find out the phonetic labels even if none were originally present.

While the separation and likelihood generation algorithms of our framework have already been improved, more attention should be paid to optimising the features and state models for maximal linguistic accuracy. The CHiME data illustrates, how some closely related words can be difficult to tell apart even under favourable conditions. Although noise robustness is a crucial aspect in practical ASR systems and our framework has

Table 3: Comparison of the recognition with three different likelihood generation methods on the test set. In addition to binary labels, OLS and PLS regression are evaluated. The best result at each SNR level and for each exemplar size is highlighted.

SNR	(dB)	9	6	3	0	-3	-6
CHiME	baseline	82.1	70.8	61.3	52.0	39.8	34.7
$T = 10$	labels	69.9	66.0	58.7	52.4	42.9	37.8
	OLS	**84.3**	**77.8**	**71.4**	**65.3**	56.4	48.6
	PLS	82.1	77.1	71.0	64.0	**57.0**	**49.3**
$T = 20$	labels	77.3	72.8	68.2	62.7	51.1	44.0
	OLS	**85.2**	**80.5**	**78.7**	**71.1**	**60.2**	**51.5**
	PLS	82.9	78.8	74.8	70.1	59.5	50.6
$T = 30$	labels	76.0	73.5	68.2	61.8	52.7	44.7
	OLS	**82.8**	**80.5**	**76.3**	**70.7**	**62.1**	**54.4**
	PLS	81.1	77.8	74.3	68.8	61.1	52.4

(a) Speaker-independent recognition

SNR	(dB)	9	6	3	0	-3	-6
CHiME	baseline	82.4	75.0	62.9	49.5	35.4	30.3
$T = 10$	labels	**91.3**	**88.3**	**85.8**	**80.8**	**71.4**	62.3
	OLS	89.8	86.8	85.0	79.7	70.1	62.7
	PLS	90.5	87.8	84.5	80.2	71.3	**63.7**
$T = 20$	labels	91.6	89.2	87.6	84.2	74.7	68.0
	OLS	91.1	**90.0**	**88.5**	**85.2**	77.6	69.2
	PLS	**91.9**	89.3	88.2	85.0	**78.6**	**69.6**
$T = 30$	labels	88.8	**88.1**	86.3	82.9	75.1	68.3
	OLS	88.8	86.0	**86.4**	**83.3**	76.1	**69.2**
	PLS	**89.1**	85.7	84.8	82.4	**77.2**	68.8

(b) Speaker-dependent recognition

shown significant advances in achieving it, the ultimate goal of maximally accurate recognition of speech itself should not be forgotten or compromised. Proper phonetic state models should be introduced instead of the current word models to avoid multiple meanings between similar features, and to make large vocabulary recognition feasible.

8. Conclusions

Exemplar-based methods were presented for recognition of speech in highly variable real world noise. The main framework and its variants were evaluated using the CHiME challenge database, which covers actual living room noise at multiple SNRs. We achieved recognition rates of over 91% at 9 dB, and close to 70% at -6 dB. Long temporal context with 200 ms exemplars, speaker-dependent speech dictionaries and adaptive noise dictionary gathering were found the best options for recognition of noisy speech.

Two separation algorithms, non-negative matrix factorisation and -deconvolution were used for determining the exemplar activations from Mel-scale spectral magnitude features. In these experiments, factorisation of overlapping windows independently from each other performed better than deconvolutive separation of whole utterances at once.

Learning the mappings from exemplar activations to state likelihoods using OLS and PLS regression was proposed. These algorithms were compared to strict binary labels acquired from forced alignment. Highest gains were seen in speaker-independent recognition. The original binary labels produced unreliable results, while mappings learnt from large training data improved the recognition rates by 4–14% (absolute). In speaker-dependent recognition the differences were small.

The results surpassed significantly both the CHiME baseline results and our previous AURORA-2 recognition rates. While the noise robustness of our system is already relatively high, parameter optimisation and better speech models would help in improving the overall recognition quality even further.

9. Acknowledgements

Antti Hurmalainen has been funded by the Academy of Finland. The research of Jort F. Gemmeke was funded by IWT-SBO project ALADIN contract 100049.

10. References

[1] S. J. Rennie, J. R. Hershey, and P. A. Olsen, “Single-channel multitalker speech recognition: Graphical modeling approaches,” *IEEE Signal Processing Magazine*, vol. 27, no. 6.

[2] P. J. Moreno, B. Raj, and R. M. Stern, “A vector taylor series approach for environment-independent speech recognition,” in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*, Atlanta, USA, 1996.

[3] B. Raj and R. M. Stern, “Missing-feature approaches in speech recognition,” *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 101–116, September 2005.

[4] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, “Exemplar-based sparse representations for noise robust automatic speech recognition,” *Accepted for publication in IEEE Transactions on Audio, Speech, and Language Processing*, 2011.

[5] T. Virtanen, J. F. Gemmeke, and A. Hurmalainen, “State-based labelling for a sparse representation of speech and its application to robust speech recognition,” in *Proceedings of INTERSPEECH*, Makuhari, Japan, 2010.

[6] G. S. V. S. Sivaram, S. K. Nemala, M. Elhilali, T. D. Tran, and H. Hermansky, “Sparse coding for speech recognition,” in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*, Dallas, USA, 2010.

[7] B. Schuller, F. Weninger, M. Wöllmer, Y. Sun, and G. Rigoll, “Non-negative matrix factorization as noise-robust feature extractor for speech recognition,” in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*, Dallas, USA, 2010.

[8] H. Christensen, J. Barker, N. Ma, and P. Green, “The CHiME corpus: a resource and a challenge for computational hearing in multisource environments,” in *Proceedings of INTERSPEECH*, Makuhari, Japan, 2010.

[9] M. Cooke, J. Barker, S. Cunningham, and X. Shao, “An audio-visual corpus for speech perception and automatic speech recognition,” *Journal of the Acoustical Society of America*, vol. 120(5), 2006.

[10] T. Virtanen, “Separation of sound sources by convolutive sparse coding,” in *Proceedings of ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing*, München, Germany, 2004.

[11] A. Hurmalainen, J. F. Gemmeke, and T. Virtanen, “Non-negative matrix deconvolution in noise robust speech recognition,” in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*, Prague, Czech Republic, 2011.

[12] P. Geladi and B. R. Kowalski, “Partial least-squares regression: a tutorial,” *Analytica Chimica Acta*, vol. 185, no. 1, 1986.

Publication P3

A. Hurmalainen and T. Virtanen, “Modelling Spectro-Temporal Dynamics in Factorisation-Based Noise-Robust Automatic Speech Recognition”, in *Proceedings of the 37th IEEE International Conference on Audio, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, 25.–30. March 2012, pp. 4113–4116.

MODELLING SPECTRO-TEMPORAL DYNAMICS IN FACTORISATION-BASED NOISE-ROBUST AUTOMATIC SPEECH RECOGNITION

Antti Hurmalainen *Tuomas Virtanen*

Tampere University of Technology, P.O. Box 553, 33101 Tampere, Finland

ABSTRACT

Non-negative spectral factorisation has been used successfully for separation of speech and noise in automatic speech recognition, both in feature-enhancing front-ends and in direct classification. In this work, we propose employing spectro-temporal 2D filters to model dynamic properties of Mel-scale spectrogram patterns in addition to static magnitude features. The results are evaluated using an exemplar-based sparse classifier on the CHiME noisy speech database. After optimisation of static features and modelling of temporal dynamics with derivative features, we achieve 87.4% average score over SNRs from 9 to -6 dB, reducing the word error rate by 28.1% from our previous static-only features.

***Index Terms*—** Automatic speech recognition, exemplar-based, spectral factorisation, noise robustness

1. INTRODUCTION

In its current state, automatic speech recognition (ASR) can achieve high phonetic classification quality in favourable conditions. However, the same cannot be said about noisy ASR. As the signal to noise ratio decreases towards zero or below, a majority of spectral features becomes corrupted, and traditional recognisers cannot match the observations to speech models reliably. Especially non-stationary noise is problematic for recogniser back-ends and difficult to counter with uniform compensation methods. Therefore detecting and removing non-speech artifacts becomes essential for noise-robust ASR.

To compare different robust ASR methods, PASCAL CHiME challenge was announced in 2010, and its results were gathered in a workshop in September 2011 [1]. As the test data includes very low SNRs, practically all challenge entries contained enhancement or separation steps for extracting real speech features from the noisy mixture [2]. Proposed approaches included beamforming, spatial uncertainty-of-observation, statistical speech-noise models and independent component analysis. Separation algorithms can thus be considered highly important for everyday ASR in general. What is less clear is how to select the algorithms and features for the task.

One significant group of separation methods consists of spectral factorisation. Due to the novelty of this branch, current work mostly focuses on modelling static spectrogram features. Nevertheless, we know that important characteristics of speech and noise can be found in spectral dynamics, that is, local changes in spectro-temporal patterns. In MFCC-based recognition, it has been found beneficial to augment the base features with *time derivatives*, also known as *delta coefficients* [3]. Another approach suggested for long temporal context modelling is using TRAP features, where the emphasis is on long term behaviour of a few spectral bands [4]. In our exemplar-based framework, spectrogram windows spanning up to 300 ms can capture a lot of temporal context [5], but some of the dynamic information is lost in the additive model. It has been suggested, that dynamics can be emphasised in factorisation-based recognition by including temporal and spectral derivatives in the feature vectors [6].

In this work, we inspect further the efficiency of derivative features on top of optimised Mel magnitudes to improve the robustness of factorisation-based recognition. The work is organised as follows. First, we introduce in Section 2 our exemplar-based factorisation framework and its recognition method known as *sparse classification* (SC). Then we describe the concept of derivative features in Section 3. The CHiME challenge data, our basic setup and feature space experiments are described in Section 4, whereafter we conclude in Section 5.

2. EXEMPLAR-BASED SPARSE CLASSIFICATION

While many separation methods are based on statistical speech and noise models, in our approach we make the models more explicit by representing the observed features as a combination of *exemplars* — spectrogram segments sampled directly from the training material or the local context [5].

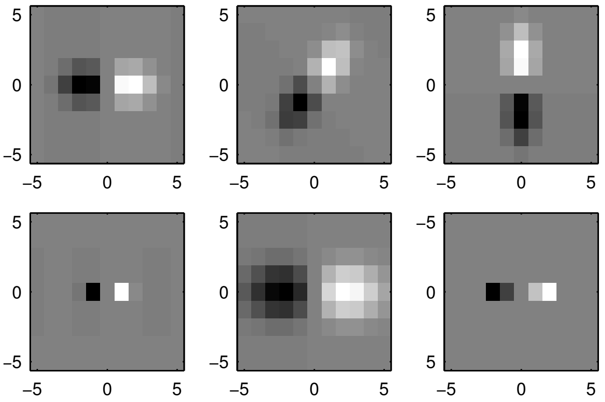
Each exemplar in our system is a $B \times T$ spectrogram matrix with B spectral bands and T consecutive frames. They are gathered to a *basis* or *dictionary*, which is used to model observed speech and noise features. Each observation window is represented as a linear combination of basis atoms. If we reshape the observation matrix to a vector $\mathbf{y}$ and each exemplar (basis atom) to a column vector $\mathbf{a}_i$, the problem becomes finding the activation weight vector $\mathbf{x}$ so that

$$\mathbf{y} \approx \sum_{i=1}^{m} \mathbf{a}_i x_i \tag{1}$$

where m is the number of exemplars in the basis. In matrix form the same equation can be given as $\mathbf{y} \approx \mathbf{A}\mathbf{x}$. Multiple observation windows can be given as parallel column vectors to solve the total activation matrix $\mathbf{X}$ ($m \times n$) for n windows at once. Finally, by assuming that basis and observation features are non-negative spectral magnitudes, and that activations should be non-negative too, finding $\mathbf{X}$ becomes a *non-negative matrix factorisation* (NMF) problem for a fixed basis. Enforcing additional sparsity on the solution ensures, that a few best fitting matches are favoured over unrealistically complex combination of multiple atoms. The iterative update rules used to find the $\mathbf{x}$ estimates are presented in [5].

To determine the utterance content from activations, each exemplar has a $Q \times T$ *label matrix*, describing the likelihood of each state $q \in [1, Q]$ over the exemplar's frames $[1, T]$. Label matrices are added together according to the corresponding exemplars' activation weights in temporal locations, where the activation was observed. This produces a $Q \times T_{\text{utt}}$ *likelihood matrix* for the whole utterance, which can be decoded using a standard Viterbi algorithm. The full procedure is described in earlier work [5, 7].

Figure 1: Spectro-temporal filters. Top row: 'Medium' length Gabor filters for temporal, diagonal and spectral direction. Bottom row: 'Short' and 'Long' Gabor filters, and length 2 HTK delta filter. Magnitudes are shown at a full greyscale range, thus not in scale.



As the decoding is based on activation weights and exemplar labels, there is no need to reconstruct the clean spectrogram or to synthesise the waveform for an external back-end. Even though spectrum or signal enhancement are also possible, in earlier work we have shown that direct classification performs better than the single-stream alternatives [5]. Multi-stream methods can improve the results significantly [8], but in this work we only use SC for simplicity and to eliminate the contribution of other components.

3. SPECTRO-TEMPORAL DERIVATIVE FEATURES

Current spectral factorisation algorithms are mostly employed in plain magnitude spaces, which model the activity in spectrogram bins, but not the dynamics over time and frequency. As in MFCC time derivatives, the NMF base features can be augmented by differential estimates. Because we are working in Mel spectrogram domain, it is possible to observe changes not only in time, but in any spectro-temporal direction by using 2-dimensional filters. The concept is similar to edge detection algorithms in image processing.

First, we construct a filter matrix in the spectro-temporal space. Then a derivative feature matrix is calculated by common 2-dimensional convolutive filtering of the static features, revealing the on- and offsets of spectrogram patterns. However, it should be noted that the differential estimates can have any sign, unlike the original non-negative magnitudes. To stay in the non-negative domain required by standard NMF algorithms, we must modify the features before factorisation.

The derivative feature matrix is reshaped to a vector, and represented by two vectors of the same size. The first contains the positive coefficients, and zeros where the vector was negative. Similarly, the second vector contains the absolute values of negative coefficients. If we denote the static features by a row vector $\mathbf{f}$ and its derivative by $d\mathbf{f}$, the augmented feature vector becomes

$$\hat{\mathbf{f}} = [\mathbf{f},\ d\mathbf{f}^{+},\ d\mathbf{f}^{-}] = [\mathbf{f},\ \max(d\mathbf{f}, \mathbf{0}),\ \max(-d\mathbf{f}, \mathbf{0})]. \quad (2)$$

If multiple derivatives are used, they are concatenated further to the vector as +/- pairs. Similar implementation was used in [6].

To learn the directions helpful for phonetic classification, we experimented with real-valued Gabor filters for multiple directions and sizes. Examples of filter matrices are shown in Figure 1, and they are described in more detail in Section 4.4.

4. EVALUATION

4.1. CHiME challenge data

The experiments were conducted using the PASCAL CHiME challenge database [1]. Its speech data consists of GRID corpus sentences, which follow a linear grammar of six word classes. The task is to recognise words belonging to the 'letter' and 'digit' classes, which contain 25 and 10 word options, respectively.

CHiME utterances are convolved with room response patterns, and mixed with household noises at six SNRs ranging from +9 to -6 dB. For training, there are 500 reverberated utterances for each of the 34 speakers, and six hours of plain background noise. The development and test sets consist of 600 utterances each, distributed between all speakers. Each set is repeated for all SNRs by mixing the utterances with different background segments containing an appropriate level of noise. All noisy utterances are presented within a long noise context as 'embedded' wave files. The development utterances are also available as 'clean' files with reverberation but no additive noise. Speaker identity is assumed to be known during recognition, while the target SNR is not.

4.2. Base setup

Our exemplar-based setup generally follows the one described in [7]. To reduce the number of parameters, we only use exemplar length of 20 frames (25 ms frame length, 10 ms shift), speaker-dependent speech bases and adaptively sampled noise bases in this work. The previous results for this setup and the GMM-based CHiME challenge baseline recogniser can be found in Table 3.

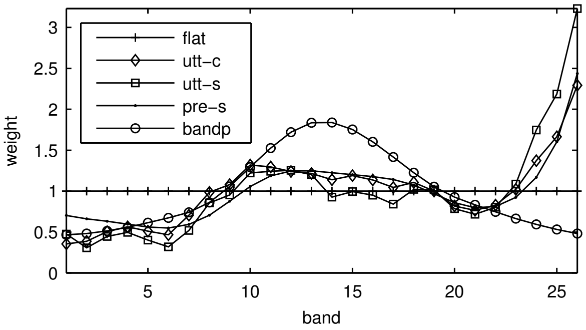
For each speaker, a speech basis is constructed by sampling 5000 exemplars from the 'clean' training speech semi-randomly. 5000 noise exemplars are also extracted for each test utterance by sampling the 'embedded' waveform files to both directions from the target utterance. In clean speech recognition, the noise basis is omitted. After converting all exemplars to Mel magnitudes and merging the speech and noise bases, a band weighting function is applied to define the contribution of each spectral band. Thereafter individual basis vectors are normalised to a Euclidean norm of 1.

Each test utterance is similarly converted into Mel magnitudes by extracting overlapping windows with a step of one frame. The band weights determined for the basis are applied to the observation as well. The observation windows are factorised to find out the activation vectors $\mathbf{x}$ as described in section 2. We initialise the activations to ones, and apply 300 rounds of an iterative update rule. The algorithm minimises the sum of estimation error (defined by KL-divergence) and a weighted L_1 penalty for non-zero activations.

As in earlier work, we used base sparsity values of 2.0 for speech and 1.7 for noise activations. However, the final sparsifying effect depends on the ratio between the penalty values and the 1-norms of basis vectors. The latter will increase by a factor of $\sqrt{R}$, if the length of 2-normed feature vectors is multiplied by R and their distribution remains similar. Therefore the $\sqrt{R}$ scaling is applied to the previously determined sparsity values, whenever the channel count, band number or derivative features change the feature vector length.

To avoid optimising for the test set, all parameter scans were performed on the development set. The 'clean' set was also used, although it does not belong to the final test set and is not included in any average values. The feature extractor was modified to use 512 FFT bins instead of the previous 256, producing small initial improvements over the earlier extraction. No changes were made to basis selection, factorisation or decoding algorithms. The learnt state mappings presented in [7] were not used in this work.

Figure 2: Mel band weighting curves for no adjustment ('flat'), online normalisation of the combined basis ('utt-c'), online speech basis normalisation ('utt-s'), precalculated normalisation from training speech ('pre-s') and bandpass filtering ('bandp').



4.3. Spectral band parameters

Before moving on to derivative features, we reoptimised the underlying static spectral magnitude space. In earlier work, we used 26 spectral bands calculated from 16 kHz signals as in the provided CHiME recogniser. The features were extracted separately for both channels, and the channel feature vectors were concatenated. These choices were re-evaluated as follows.

4.3.1. Band weighting

The Mel-scale distribution of speech and noise features is considerably uneven across bands. We can reweight the bands for two different goals; either to flatten the distribution for equal contribution of each band, or alternatively to emphasise certain bands for maximal classification quality. While the highpass filter commonly employed in MFCC extraction can improve clean speech recognition, we have found it too drastic for robust factorisation algorithms. Instead, five different weighting methods were tested:

1. No weighting ('flat')
2. Normalisation of the combined utterance basis bands ('utt-c')
3. Normalisation calculated from the speech basis only ('utt-s')
4. Precalculated normalisation of training speech bands ('pre-s')
5. Experimental bandpass filtering ('bandp')

Method 2 is our previous approach and depends on the adaptive noise basis of each utterance. Method 3 only depends on the current speech basis, that is, speaker identity. Methods 4 and 5 both produce fixed weighting, which simplifies the later steps. The bandpass weighting was included as an example of filter types, which emphasise the speech formant area and mostly discard frequencies over 4 kHz. All weighting methods are illustrated in Figure 2. For non-fixed weightings, means over all development data are shown.

The results are summarised in the first part of Table 1. We observe that 'do nothing' and online-computed speech weighting fare worse at certain SNRs than the other methods, which are approximately tied. Interestingly, the fixed weightings produce similar average rates, while bandpass filtering favour the clean end and precalculated speech normalisation the noisy one. The latter was chosen for further experiments due to its robustness, normalising effect and fixed shape. The differences between diverse weighting methods were generally small.

Table 1: Development set results for different spectral band parameter combinations. The format of experiment names is [band number] / [mono | stereo] / [weighting type].

SNR (dB)	clean	9	6	3	0	-3	-6	avg
26/s/flat	92.7	90.6	90.5	88.3	83.5	79.1	71.8	84.0
26/s/utt-c	93.7	91.8	91.8	89.8	83.5	78.5	72.2	84.6
26/s/utt-s	93.7	92.0	91.6	89.3	83.3	77.4	70.4	84.0
26/s/pre-s	93.6	91.4	90.8	89.3	84.7	78.9	72.7	84.6
26/s/bandp	93.7	92.0	91.7	89.8	83.8	78.8	71.8	84.6
26/m/pre-s	93.3	92.1	91.4	89.3	83.9	78.7	71.9	84.5
26/m/bandp	93.7	91.8	91.7	89.6	83.8	79.5	71.6	84.7
40/m/pre-s	93.6	92.3	91.6	89.8	85.0	79.7	72.7	85.2

4.3.2. Channel count

In our original parametrisation, binaural features were kept in separate entries of the feature vector, retaining some of the spatial information of the sound sources. To study whether it plays any role in recognition quality, the development set was also factorised using mono features by averaging the Mel magnitudes of channels. Apart from adjusting the sparsity value due to vector length halving, no other changes were made. Two fixed weighting curves, precalculated normalisation and bandpass filtering, were tested.

As can be seen from the results in Table 1 (rows 4–7), the accuracy of mono and stereo features is highly similar. Because mono features reduce the vector length and consequently computing costs by a half, they were used for further experiments.

4.3.3. Spectral band number

One fundamental question regarding feature selection is the number of Mel bands. To inspect this briefly, the band count was increased from 26 to 40. The results are shown on the last row of Table 1. We observe some $\sim$1% improvements and no decrements, suggesting that the gains may be worth the increased computational costs. While the next section was still evaluated using the original 26 bands, the final evaluation was performed on both values.

4.4. Spectro-temporal filters

After determining efficient base features, we tested three combinations of spectro-temporal Gabor filters: only temporal (forward and backwards), cardinal directions (temporal and spectral), and diagonal filters (45° angles). The prototype filter matrix was defined by

$$g(x,y) = \exp(-(\frac{x^2 + (\gamma y)^2}{2\sigma^2}) \sin(\frac{2\pi x}{\lambda}), \quad x, y \in [-5, 5] \quad (3)$$

with ellipticity γ set to 3, Gaussian envelope width factor σ to 2, and wavelength λ to 9, producing approximately one full sinusoid cycle. The prototype filter and two of its rotations are shown on the first row of Figure 1. The absolute sum of filter coefficients was set to 0.6 for each half of the filter. The results of augmenting directional filters to fixed-norm weighted mono features can be seen in the first part of Table 2. We notice that temporal direction improves the recognition rates, while including any of the spectral directions does not.

Settling for primarily temporal filtering, we tested the Gabor filter with its size increased and decreased by 50%, and in addition the delta filter employed by HTK using the default window length of 2 frames to both directions [3]. All were normalised to a 0.6

Table 2: Development set results for 2D filtering. Filter type is either Gabor [short | medium | long] in [temporal | cardinal | diagonal] directions, or HTK delta.

SNR (dB)	clean	9	6	3	0	-3	-6	avg
G/med/temp	92.8	92.0	91.7	89.5	83.9	80.3	73.2	85.1
G/med/card	92.9	91.8	90.3	89.3	83.8	78.1	71.7	84.1
G/med/diag	92.8	91.1	90.6	88.3	82.7	76.3	69.5	83.1
G/short/temp	93.3	92.2	92.2	90.3	85.6	81.4	73.5	85.9
G/long/temp	92.3	91.0	89.8	88.3	82.2	77.4	70.5	83.2
HTK delta	93.4	92.4	91.8	90.3	85.1	82.1	74.1	86.0

coefficient sum per side. The filters are shown on row 2 of Figure 1, and the results in the second part of Table 2. The best results were achieved using the shorter filters with little or no cross-band bleeding. The clean speech recognition rate does not improve over unfiltered base features, but the robustness against heavy noise increases. Changing the filter weight (not shown) did not produce any significant improvements.

4.5. Final test set evaluation

After optimisations, the test set was evaluated using the following parameter combination; mono features, precalculated speech-normalising band weights, and length 2 temporal delta filtering at weight 0.6. Both 26 and 40 spectral bands were used for determining their quality-cost tradeoff. The results are listed in Table 3. We notice significant improvements at each SNR in comparison to our earlier results. The word error rate is reduced by 13.9–32.8% at different SNRs, and the total error rate by up to 28.1%. Using 40 bands produces a large boost at -6 dB and modest gains elsewhere.

While the overall rates do not match the state-of-the-art results achieved in the CHiME workshop, where the best average score was 91.65% [9], it should be noted that the current highest ranking methods are relatively complex combinations of multiple techniques, whereas the approach presented here is a single stream classifier. Preliminary experiments suggests, that using sparse classification with complementary methods in multi-stream recognition can indeed achieve over 90% average recognition rate on the CHiME data already with the earlier, unoptimised features [8].

5. CONCLUSIONS

We studied alternative parametrisations of Mel features and their derivatives for factorisation-based speech recognition using CHiME challenge data and an exemplar-based sparse classifier.

First, we found out that the recognition algorithm is not particularly sensitive to band weighting, although some normalisation will improve the results over do-nothing. Mono features were found as effective as stereo for this data, allowing a 50% reduction in computational costs. Increasing the spectral band number from original 26 to 40 improved the results slightly.

Spectro-temporal filters were applied to the basis and observation features to model dynamic behaviour. Including temporal delta information produced significant improvements, while edge detection in spectral directions was found detrimental. The best temporal filters were relatively short with roughly 20ms temporal context to both directions, and no cross-band bleeding.

All in all, our feature space optimisation yielded 28.1% reduction in the total word error rate over all noisy conditions. Clean speech recognition rate remained at approximately 93–94%, which illustrates the difficulty of short word classification when no clues of word identity can be found from the neighbouring word context.

Table 3: Test set scores (%) for the CHiME baseline GMM recogniser, our previous SC features, and optimised features with their relative word error rate reductions (%) from the earlier results.

SNR (dB)	9	6	3	0	-3	-6	avg
GMM baseline	82.4	75.0	62.9	49.5	35.4	30.3	55.9
original SC, B=26	91.6	89.2	87.6	84.2	74.7	68.0	82.5
optimised SC, B=26	92.8	91.3	89.8	87.9	82.2	75.8	86.6
WER reduction	13.9	19.9	17.5	23.7	29.6	24.5	23.4
optimised SC, B=40	92.9	91.8	90.1	88.4	82.9	78.5	87.4
WER reduction	15.9	24.6	20.1	26.8	32.6	32.8	28.1

While the presented work was tested on the exemplar-based recogniser, it can be generalised to other algorithms based on non-negative spectral factorisation. The improved separation quality should prove useful both for feature-enhancing front-ends and for direct classifiers in standalone or combined recognition.

6. REFERENCES

[1] H. Christensen, J. Barker, N. Ma, and P. Green, "The CHiME Corpus: a Resource and a Challenge for Computational Hearing in Multisource Environments," in *Proc. INTERSPEECH*, Makuhari, Japan, 2010, pp. 1918–1921.

[2] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "Overview of the PASCAL CHiME Speech Separation and Recognition Challenge," in *Proc. Machine Listening in Multisource Environments (CHiME 2011), satellite workshop of INTERSPEECH 2011*, Florence, Italy, 2011.

[3] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book Version 3.3*, Cambridge University Press, 2005.

[4] H. Hermansky and S. Sharma, "TRAPs – Classifiers of Temporal Patterns," in *Proc. ICSLP*, Sydney, Australia, 1998, pp. 1003–1006.

[5] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based Sparse Representations for Noise Robust Automatic Speech Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2067–2080, 2011.

[6] M. Van Segbroeck and H. Van hamme, "Unsupervised Learning of Time-Frequency Patches as a Noise-Robust Representation of Speech," *Speech Communication*, vol. 51, no. 11, pp. 1124–1138, 2009.

[7] A. Hurmalainen, K. Mahkonen, J. F. Gemmeke, and T. Virtanen, "Exemplar-based Recognition of Speech in Highly Variable Noise," in *Proc. CHiME workshop*, Florence, Italy, 2011, pp. 1–5.

[8] F. Weninger, M. Wöllmer, J. Geiger, B. Schuller, J. F. Gemmeke, A. Hurmalainen, T. Virtanen, and G. Rigoll, "Non-negative Matrix Factorization for Highly Noise-Robust ASR: To Enhance or to Recognize?," in *Proc. ICASSP*, Kyoto, Japan, 2012.

[9] M. Delcroix et al., "Speech Recognition in the Presence of Highly Non-stationary Noise Based on Spatial, Spectral and Temporal Speech/Noise Modeling Combined with Dynamic Variance Adaptation," in *Proc. CHiME workshop*, Florence, Italy, 2011, pp. 12–17.

Publication P4

A. Hurmalainen, J. F. Gemmeke, and T. Virtanen, “Detection, Separation and Recognition of Speech From Continuous Signals Using Spectral Factorisation”, in *Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, Bucharest, Romania, 27.–31. August 2012, pp. 2649–2653.

DETECTION, SEPARATION AND RECOGNITION OF SPEECH FROM CONTINUOUS SIGNALS USING SPECTRAL FACTORISATION

Antti Hurmalainen[⋆] *Jort F. Gemmeke*[†] *Tuomas Virtanen*[⋆]

[⋆] Tampere University of Technology, P.O. Box 553, FI-33101 Tampere, Finland
[†] KU Leuven, Department ESAT-PSI, Kasteelpark Arenberg 10, 3001 Heverlee, Belgium

ABSTRACT

In real world speech processing, the signals are often continuous and consist of momentary segments of speech over non-stationary background noise. It has been demonstrated that spectral factorisation using multi-frame atoms can be successfully employed to separate and recognise speech in adverse conditions. While in previous work full knowledge of utterance endpointing and speaker identity was used for noise modelling and speech recognition, this study proposes spectral factorisation and sparse classification techniques to detect, identify, separate and recognise speech from a continuous noisy input. Speech models are trained beforehand, but noise models are acquired adaptively from the input by using voice activity detection without prior knowledge of noise-only locations. The results are evaluated on the CHiME corpus, containing utterances from 34 speakers over highly non-stationary multi-source noise.

Index Terms— Spectral factorization, speech recognition, speaker recognition, voice activity detection, speech separation

1. INTRODUCTION

Applying automatic speech recognition (ASR) in noisy environments introduces several new challenges not present in clean conditions. A fundamental problem is corruption of speech features by additive noise, which may not match to noise observed during model training. In previous work, high separation quality has been achieved by applying spectral factorisation that decomposes a noisy input spectrogram into activations of multi-frame speech and noise *atoms*, which can be acquired from training material or from the local context [1, 2, 3, 4]. We have shown that a method known as *sparse classification* (SC), which determines the phonetic content directly from the weights of activated speech atoms, can produce speech recognition results comparable to source separation followed by conventional back-end recognition [1, 5].

In previous experiments with noise atoms sampled from the neighbourhood of noisy utterances, we have used annotated speech endpointing to sample from segments known to consist of only noise. In real world applications, such information cannot be assumed to be available, thus speech activity must be estimated. In other speech recognition methods, voice activity detection (VAD) has been employed to detect speech and noise segments and to update the noise model [6].

In this work, we propose the use of SC-based methods for detecting the target utterances from mixtures containing high noise levels and occasionally overlapping non-target speech. The same framework is used for noise model updating and subsequent source separation. Speech models are acquired beforehand from training material, whereas noise models are adapted from the context.

Another topic of interest is the use of speaker-dependent speech recognition to obtain better results in both clean and noisy environments. However, the true speaker identity may not be known during recognition. We propose SC for determining the speaker identity from continuous noisy mixtures, whereafter source separation and speech recognition is carried out with speaker-dependent speech models.

The work is organised as follows: Section 2 introduces the main concepts of factorisation-based speech separation and recognition. In Section 3 we present the framework for processing continuous audio, detecting speech locations, and updating the noise model. In Section 4 we apply the algorithms to CHiME data, consisting of utterances from 34 speakers over continuous, highly non-stationary background noise. Finally, in Section 5 we draw the conclusions.

2. FACTORISATION-BASED SPEECH SEPARATION AND RECOGNITION

The methods presented here are based on representing an observed sound mixture as a linear sum of speech and noise *atoms*, each belonging to a single speaker or to background noise. The features consist of Mel scale spectral magnitudes, computed in 25 ms *frames* with a 10 ms shift. The atoms are $B \times T$ spectrogram segments, where B is the number of Mel bands and T is the number of consecutive frames in an atom. Speech and noise atoms form a *dictionary* (or *basis*). By assuming that magnitudes of multiple sources are approximately additive in the Mel-spectral domain, factorisation becomes a problem of finding non-negative *activation weights* x_l for each atom index $l \in [1, L]$ in the system, together denoted as an *activation vector* $\mathbf{x}$.

The research of Jort F. Gemmeke was supported by IWT-SBO project ALADIN contract 100049.

2.1. Convolutive spectral factorisation

An observation spectrogram $\mathbf{Y}$ ($B \times F$), where the number of frames F is larger than atom duration T, is factorised using convolutive temporal modelling and joint spectrogram estimation with overlapping segments [7]. We find the $L \times W$ *activation matrix* $\mathbf{X}$, consisting of an activation vector for all W *window indices* in the observation. We only consider windows fitting completely within the observation spectrogram. Thereby the activations of the final window index W takes place at time $F - T + 1$. $\mathbf{X}$ is obtained by optimising the estimated observation spectrogram $\mathbf{\Psi}$, modelled convolutively as

$$\mathbf{\Psi} = \sum_{t=1}^{T} \mathbf{A}_t \overset{\rightarrow(t-1)}{\mathbf{X}}. \tag{1}$$

Each $\mathbf{A}_t$ ($t \in [1, T]$) is a $B \times L$ matrix, containing frame t of every $B \times T$ atom in the dictionary. Operator $\rightarrow$ shifts the columns of $\mathbf{X}$ right within a $L \times F$ zero-padded matrix by $t - 1$ columns. The cost function to be minimised consists of Kullback-Leibler divergence between $\mathbf{Y}$ and $\mathbf{\Psi}$, and the sum of $\mathbf{X}$ entries weighted element-wise by a sparsity penalty matrix. The exact cost functions and iterative update rules used in our convolutive factorisation are described in [2, 5].

2.2. Source separation and sparse classification

The activation matrix can be used for source separation. Two spectrogram estimates are derived from Equation 1; a noisy speech reconstruction $\mathbf{\Psi}$ obtained by using both speech and noise atoms, and a clean speech estimate $\mathbf{\Psi}_s$ obtained by only using speech atoms and activations. The element-wise speech-to-total ratio $\mathbf{\Psi}_s/\mathbf{\Psi}$ is converted back to discrete Fourier frequency scale by multiplication from the left by a pseudoinverse of the Mel filterbank matrix, and acts as a time-varying filter for the original mixture spectrogram. It is then used to estimate speech-only features and further to synthesise separated time-domain signals [5].

To determine the speaker identity and phonetic content from speech atom activations, each speaker-dependent speech atom is associated with a $Q \times T$ *label matrix* $\mathbf{B}$. It represents the presence of each phonetic state $q \in [1, Q]$ over the atom's frame indices [1, 5]. These atom-state labels are used to calculate a $Q \times F$ *likelihood matrix*, representing phonetic state likelihoods over the whole duration of the observation. The likelihood matrix is calculated by applying Equation 1, with state label matrices $\mathbf{B}$ taking the place of the atom spectrograms. The method is known as *sparse classification*. In previous work we have used it for speech decoding [1, 2, 5]. Here the state likelihood information is used for voice activity detection and speaker identification.

3. PROPOSED SYSTEM FOR PROCESSING CONTINUOUS AUDIO

In the proposed system, continuous input audio is processed gradually using convolutive spectral factorisation, a fixed multi-speaker speech basis obtained in the training stage, and an adaptively updated noise basis. As the factorisation advances within the signal, speech activation weights and state mapping matrices are used to construct estimates of the presence of phonetic states for each speaker individually. The speaker-dependent state information is used for two purposes, speech locating and speaker identification.

3.1. Voice activity detection

We perform initial factorisation in 750-frame (7.5 s) spectrogram *blocks*. An extended Hann window function, consisting of 250 frames of fade-in, 250 frames of flat top and 250 frames of fade-out is applied to each block spectrogram. 2/3 overlap is present between blocks, so that each frame of the input is included in exactly one flat middle section. Blocks are factorised consecutively using the convolutive model described in Section 2.1 with a multi-speaker speech basis (Section 4.2) and an adaptive noise basis (Section 3.2).

Speech activations are converted into phonetic state likelihoods by using mapping matrices $\mathbf{B}$ and overlap-added over blocks. Using the initial state likelihood estimates and word-dependent *VAD weight functions* over time, a total VAD level estimate is derived for each input frame. Each word is assigned a specific weight profile over time, spanning up to 30 frames to both directions from the original frame location for temporal smoothing and utterance modelling. Based on the task grammar, the shape of weight functions depends on the role of each word in a sentence: the functions corresponding to the first and last word classes in a sentence are given negative weight before and after them, respectively. This emphasises the contrast in VAD level between target speech and its surroundings, helping to isolate test utterances from noise and non-test speech. An example of weight functions that were used in the simulations is shown in Figure 1. Word activity sums are convolved with their respective weight functions, and then summed together for the total VAD weight.

Speech-noise classification is performed using the total VAD weight over frames and on/off threshold values determined from development data. In addition, constraints can be set on the utterance duration to select candidates matching to the expected temporal profile of utterances.

3.2. Noise basis acquisition

Areas flagged as noise are sampled directly into noise atoms with a $T/2$ overlap between consecutive atoms. A threshold value is used on the spectrogram magnitude sum of segments to only store atoms with significant noise events. A noise dictionary is maintained, starting empty and acquiring new content up to a defined maximum capacity. Each noise dictionary atom is given a *significance weight*, increasing according to its activation weight in factorisation and decaying exponentially over time. Whenever newly introduced noise atoms would exceed the dictionary size, the least significant existing atoms are discarded. The latest dictionary is always used for factorisation.

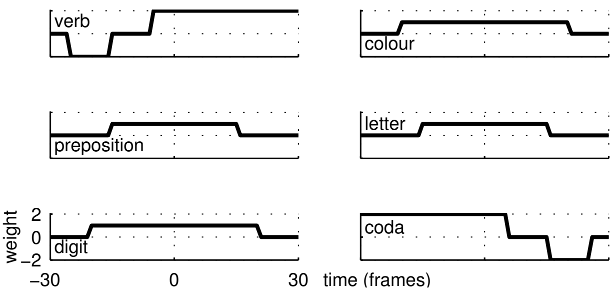


Fig. 1. VAD weight functions for each CHiME word class.

3.3. Speaker recognition

As we use a multi-speaker basis with knowledge of the speaker identity of each speech atom, a likelihood matrix can be generated for each speaker individually. For a span of frames marked as speech, we find the maximum sum of speaker-dependent state content to identify the most likely speaker. The identification result, in turn, is used for another, local factorisation pass so that only the chosen speaker's speech basis is included. By narrowing down the speech basis, the system becomes more sensitive to the chosen identity and may be able to pick the correct phonetic content even from mixtures containing other speakers. In separation-based speech recognition, the identity estimate is also used for selecting the speaker-dependent GMM model in the back-end.

4. EXPERIMENTS

4.1. CHiME data

The experiments were conducted on CHiME data, consisting of GRID command utterances mixed over highly non-stationary family household noises with simulated room reverberation response matching the noise [8]. The target utterances are from 34 different speakers and follow a linear six-class *verb-colour-preposition-letter-digit-coda* grammar ("set white in H 7 please"). A default language model is provided for recognition, employing 250 sub-word states for the 51-word vocabulary. For each speaker, there are 500 training utterances with reverberation but no additive noise. Development and test sets consist of a total of 600 utterances from all speakers together, repeated at multiple SNR levels. The noises contain a large variety of everyday sound events including appliances, impacts, music and also spontaneous speech from non-target speakers.

For this work, we use the continuous, 'embedded' CHiME sequences. In the test set, there are 16 *sessions* ranging from 27 to 87 minutes. The 600 test utterances are spread over the sessions at SNRs ranging from +18 to -6 dB at 3 dB intervals. SNRs from +9 to -6 dB belong to the official scoring set. The locations of speech in sessions are chosen in such a way that the target SNR is achieved by direct mixing without scaling. Therefore it is common for one loud segment of background noise to contain several low-SNR test utterances in succession. Conversely, there are also long noise-only sequences between the test utterances.

All 16 kHz CHiME audio was converted into $B = 40$ band Mel-scale magnitude spectrograms with 25 ms frame length and 10 ms frame shift, and equalised using a frequency band weighting curve derived from speech training material. For spectral processing, the magnitudes of left and right channels were averaged to form monaural spectrogram features.

4.2. Bases and labelling

A speech basis was created for every speaker by employing forced alignment data acquired from the CHiME HTK models. Based on the 250 sub-word phonetic states, each state in turn was modelled by placing its corresponding word instances from 300 training utterances in a $B \times T$ spectrogram window with the target state in the middle [5]. A median was taken over the instances within each time-frequency point, creating a characteristic template of the state spectrum and its typical neighbourhood. Atom length T was set to 25 (265 ms), which is enough to capture short words in their entirety, and partial content of longer words, together modelling slight variations in the pace of pronunciation. All in all, the 250 atoms of 34 speakers formed a 8500-atom speech basis. The remaining 200 training utterances from all speakers were combined and factorised using the full speech basis to learn the activation-state mapping matrices $\mathbf{B}$ with ordinary least squares regression as described in [2, 5].

An adaptive noise basis was maintained as described in Section 3.2. We used a maximum capacity of 500 atoms for sampled noise. In addition, 15 atoms were initialised randomly and updated during iteration to model unseen noise events, e.g. when the adaptive basis was empty [5]. The maximum number of atoms used in block factorisation was 9015 (8500 speech, 500 sampled noise, 15 on-line updated noise).

4.3. VAD accuracy

The VAD algorithm described in Section 3.1 was used to find utterances from CHiME sessions. A VAD weight function was given for each word class in CHiME grammar to reflect the expected speech activity profile in its neighbourhood. The functions are shown in Figure 1. On/off thresholds for total VAD level were acquired from development data and set to favour false positives over missed true utterances. To reflect the duration of CHiME utterances, a minimum length requirement of 80 frames was set for speech segments, and after 180 frames from the start of a segment it was ended as soon as the silence threshold was reached. Between these limits, temporary gaps of up to 60 frames were allowed to model short pauses in speech. Because the CHiME ground truth annotations occasionally contain excess silence, an utterance was ruled as being found in a segment for scoring if at least 40% of its duration was flagged as speech by VAD.

Speech detection results are listed in Table 1. Of the 5400 test utterances (600 for each SNR level), 5331 (98.7%) were detected successfully. 5090 were also assigned correctly to single segments, whereas 241 appeared in segments where

Table 1. Voice activity detection results: 600 utterances at 9 SNR levels, all in all 5400 utterances, exist within the continuous test sessions. 5939 speech segments were detected.

Found speech segments	True utterances
726 false positives 5090 containing 1 utterance 120 containing 2 utterances 3 containing 3 utterances	65 false negatives (misses) 5331 found in 1 segment 4 split between 2 segments
Total: 5939	Total: 5400

two or more consecutive utterances got merged. In a few cases, an utterance was split between two found segments. 726 false positives — segments with no target utterances — were also found. These mostly consisted of other speech found in CHiME background noise.

In a completely realistic scenario, the detected speech segments should be identified and recognised by themselves. In these experiments we used found segments for VAD quality evaluation and noise modelling, but annotated endpointing for speaker identification and speech recognition. The reason for this choice is that the default CHiME language model assumes tightly cropped, single utterances as its input. Passing VAD-based segments with possible silence and merged utterances would result in unpredictable back-end behaviour and problems in comparing the scores with prior ASR methods.

4.4. Speaker identification

The results for speaker identification are listed by SNR in Table 2. The identification rates of 12–18 dB utterances were between 99–100%. We notice that above 0 dB, misclassifications are rare. From 0 dB downwards, the utterances may — and often do — contain equally loud speech from non-target speakers, which may cause the maximum activity classifier to select an identity matching to the non-target speech instead of the true speaker. Misclassification of target speech to another similar sounding speaker may also take place due to corruption of spectral features. For enhancement and sparse classification, the latter kind of errors are still tolerable, whereas the former are often unrecoverable.

4.5. Speech separation and recognition

Enhanced utterances were cropped from the full session signals separated during multi-speaker block processing. Real utterance locations were also re-factorised using a single-speaker basis of both true and estimated speaker identity in turn. The latest sampled noise basis and $\lceil F/T \rceil$ on-line updated noise atoms were used in the second, local factorisation pass. Enhanced test signals, generated as described in Section 2.2, were stored for GMM-based speech recognition and measurement of signal-to-distortion ratio (SDR) of enhanced utterances in comparison to clean test files.

For enhancement-based recognition, we used the CHiME language model and multi-condition (MC) trained speaker-dependent GMMs as in [3, 5]. Models were not retrained for enhanced signals. SDR was calculated as

Table 2. Speaker identification scores (%) on the CHiME test set over SNRs. SC-based maximum state sum is used to determine the most likely identity among 34 speakers.

SNR	9 dB	6 dB	3 dB	0 dB	-3 dB	-6 dB	avg
Correct	99.2	98.7	97.2	91.3	85.0	74.5	91.0

$$SDR_{dB} = 10 \log_{10} \frac{\sum_n s(n)^2}{\sum_n (\hat{s}(n) - s(n))^2}, \qquad (2)$$

where $s(n)$ is the clean reference signal and $\hat{s}(n)$ is the noisy or enhanced signal over sample index n [9]. Because CHiME annotations do not match perfectly to the isolated files, signals were aligned with maximum cross-correlation before measurement. Both in recognition and SDR measurement, left and right channels were averaged to form monaural signals.

Results for speech recognition are shown in Table 3. The first half displays baseline scores for the clean-trained CHiME reference models, the MC-trained models without any enhancement, and our previous results using a 250-atom sampled noise dictionary exploiting full knowledge of noise-only segments and the same speech bases as in this work [5]. In the second half, recognition results are shown for the new VAD-based noise modelling. Four different combinations are used for the choice of speech dictionaries in factorisation and for speaker-dependent GMM models used in the back-end.

The scores generally decrease as endpointing and identity information is lost, but even in the worst case where estimated identity is used for all parts, the new results surpass unenhanced, known-identity recognition by a wide margin. Interestingly, enhancement using all speakers' bases is on average better than only using the true identity. One possible explanation is that using all bases simultaneously allows wider phonetic variation, even though not all atoms belong to the target speaker. The degradation from losing identity information in separation and GMM selection reflects the misclassification rates over SNRs seen in Table 2. The largest decrements take place in the noisy end, but overall only 2.6% (absolute) loss is observed in average accuracy when true identity is wholly replaced by an estimate.

Results for SDR measurement are shown in Table 4. The first rows show SDRs for unenhanced utterances and enhancement with the earlier 250-atom informed noise modelling. Note that the nominal CHiME SNRs do not match the measured, unenhanced SDRs due to different weighting. In the second part, results for the new, self-adapting noise model are shown. Curiously, our new model produces superior separation quality, which does not translate to better ASR rates. We can speculate that the proposed noise model with long memory and adaptive atoms is able to remove more major noise events than the strictly local, informed model. Meanwhile, it may also remove crucial speech information, thus reducing

Table 3. Enhancement-based speech recognition scores (%) on the CHiME test set over different SNRs. First part shows unenhanced baseline scores for standard CHiME models and multi-condition (MC) trained models, and the latter with informed 250-atom noise modelling. The second part uses new, self-adapting noise models. Row labels denote the speech bases used for enhancement (all/true/estimated), and the speaker model used for GMM evaluation (true/estimated).

SNR	9 dB	6 dB	3 dB	0 dB	-3 dB	-6 dB	avg
Baseline scores and informed noise modelling							
CHiME	82.4	75.0	62.9	49.5	35.4	30.3	55.9
MC, none	91.3	86.8	81.7	72.8	61.1	54.5	74.7
MC, inform.	93.0	91.2	90.0	85.2	79.0	72.9	85.2
Self-adapting noise, enhancement + MC recognition							
All/true	92.8	89.8	87.8	84.4	75.5	73.9	84.1
True/true	91.6	88.8	88.2	83.9	76.9	68.9	83.0
Est./true	91.4	88.8	87.8	82.6	73.8	64.4	81.5
Est./est.	91.4	88.6	87.1	81.0	72.3	62.2	80.4

Table 4. Measured signal-to-distortion ratios (dB) for unenhanced and enhanced CHiME test utterances over nominal mixing SNRs.

SNR	9 dB	6 dB	3 dB	0 dB	-3 dB	-6 dB	avg
Unenhanced signals and informed noise modelling							
Unenhanced	3.7	2.5	0.3	-1.9	-4.8	-7.0	-1.2
Informed	4.4	4.1	3.8	3.5	3.1	2.7	3.6
Self-adapting noise, all/true/estimated identity							
All	8.6	7.8	6.8	5.9	4.7	3.9	6.3
True	6.9	6.4	6.0	5.5	4.9	4.4	5.7
Estimated	6.9	6.4	6.0	5.4	4.6	4.0	5.6

the final ASR rate. Another noteworthy observation is that the compact single-speaker speech models introduce more distortions in the clean end than using all speakers' bases, but in the noisy end they manage to separate target speech better.

5. CONCLUSIONS

Spectral factorisation based methods were presented for solving three problems; voice activity detection, speaker identification, and speech separation/recognition from a continuous input. Results were evaluated using CHiME data, containing 34 speakers and household noise at SNRs from 9 to -6 dB.

98.7% of target utterances were found by estimating voice activity from speech atom activations and state labels. False positives generally consisted of non-target speech present in CHiME noise. Non-speech segments were used to update the noise model in continuous factorisation, thereby making the model completely independent of noise training data.

Activation weights of a multi-speaker basis were used to determine speaker identity among the 34 candidates. An average identification rate of 91.0% was achieved over all SNRs. Thereafter utterances were separated for GMM-based speech recognition. The new, self-adapting noise model yielded higher signal-to-distortion ratios than earlier, informed noise modelling. However, speech recognition rates decreased slightly when speaker identity was estimated. Approximately 80% average scores were still achieved after bypassing all information on speaker identity and noise locations.

The results as a whole demonstrate, how spectrogram factorisation and sparse classification can be used for several subtasks in noise-robust speech separation and recognition. We eventually hope to extend the presented work into a complete large vocabulary continuous speech recognition framework based on SC techniques.

6. REFERENCES

[1] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based Sparse Representations for Noise Robust Automatic Speech Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2067–2080, 2011.

[2] A. Hurmalainen, K. Mahkonen, J. F. Gemmeke, and T. Virtanen, "Exemplar-based Recognition of Speech in Highly Variable Noise," in *Proc. CHiME workshop*, Florence, Italy, 2011, pp. 1–5.

[3] F. Weninger, J. Geiger, M. Wöllmer, B. Schuller, and G. Rigoll, "The Munich 2011 CHiME Challenge Contribution: NMF-BLSTM Speech Enhancement and Recognition for Reverberated Multisource Environments," in *Proc. CHiME workshop*, Florence, Italy, 2011, pp. 24–29.

[4] R. Vipperla, S. Bozonnet, D. Wang, and N. Evans, "Robust Speech Recognition in Multi-Source Noise Environments using Convolutive Non-Negative Matrix Factorization," in *Proc. CHiME workshop*, Florence, Italy, 2011, pp. 74–79.

[5] A. Hurmalainen, J. F. Gemmeke, and T. Virtanen, "Modelling Non-stationary Noise with Spectral Factorisation in Automatic Speech Recognition," *submitted work*, 2012.

[6] K. Demuynck, X. Zhang, D. Van Compernolle, and H. Van hamme, "Feature versus Model Based Noise Robustness," in *Proc. INTERSPEECH*, Florence, Italy, 2011, pp. 721–724.

[7] P. Smaragdis, "Convolutive Speech Bases and their Application to Supervised Speech Separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 1–14, 2007.

[8] H. Christensen, J. Barker, N. Ma, and P. Green, "The CHiME Corpus: a Resource and a Challenge for Computational Hearing in Multisource Environments," in *Proc. INTERSPEECH*, Makuhari, Japan, 2010, pp. 1918–1921.

[9] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. P. Rosca, "First Stereo Audio Source Separation Evaluation Campaign: Data, Algorithms and Results," in *Proc. ICA*, 2007, pp. 552–559.

Publication P5

A. Hurmalainen, R. Saeidi, and T. Virtanen, “Group Sparsity for Speaker Identity Discrimination in Factorisation-based Speech Recognition”, in *Proceedings of the 13th INTERSPEECH*, Portland, Oregon, USA, 9.–13. September 2012, pp. 2138–2141.

ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Group Sparsity for Speaker Identity Discrimination in Factorisation-based Speech Recognition

Antti Hurmalainen[1], *Rahim Saeidi*[2], *Tuomas Virtanen*[1]

[1]Department of Signal Processing, Tampere University of Technology, Tampere, Finland
[2]Centre for Language and Speech Technology, Radboud University Nijmegen, The Netherlands
antti.hurmalainen@tut.fi, rahim.saeidi@let.ru.nl, tuomas.virtanen@tut.fi

Abstract

Spectrogram factorisation using a dictionary of spectro-temporal atoms has been successfully employed to separate a mixed audio signal into its source components. When atoms from multiple sources are included in a combined dictionary, the relative weights of activated atoms reveal likely sources as well as the content of each source. Enforcing sparsity on the activation weights produces solutions, where only a small number of atoms are active at a time. In this paper we propose using group sparsity to restrict simultaneous activation of sources, allowing us to discover the identity of an unknown speaker from multiple candidates, and further to recognise the phonetic content more reliably with a narrowed down subset of atoms belonging to the most likely speakers. An evaluation on the CHiME corpus shows that the use of group sparsity improves the results of noise robust speaker identification and speech recognition using speaker-dependent models.

Index Terms: group sparsity, speech recognition, speaker identification, spectrogram factorization

1. Introduction

In several studies it has been reported, how spectrogram factorisation using a dictionary of atoms has produced strong results in separating multiple non-stationary sources from mixed observations [1, 2, 3]. However, a common assumption is that only certain sources are active in the mixture — for example, one known speaker over background noise, or two known speakers. Under this assumption, only the relevant dictionaries are chosen for the factorisation task, thus reducing problem complexity and confusion with sources not present in the mixture. In reality, the set of potential sources may be significantly larger than the number of sources active in the mixture, and the identities of active sources may not be known beforehand. There is ongoing research on multi-talker tasks with non-negative matrix factorisation (NMF) given as one option, but thus far the performance of its basic form has not been found satisfactory [4].

It has been shown that activations of dictionary atoms acquired via NMF can act as evidence for both the speaker identity and the phonetic content of speech [3, 5, 6]. Enforcing sparsity on the activations improves the classification results [5]. Therefore the method is referred to as *sparse classification* (SC). A straightforward sparsity constraint is to penalise all non-zero activation weights by adding a weighted L_1 norm of all activations to the cost function to be minimised. The problem of this approach is that the acquired solution may contain atoms from any number of sources as long as the distribution of individual atoms is sparse. The same spectral features may carry a different meaning if taken from another source, thereby harming the classification outcome. If we expect only a limited number of sources to be active at a time, it would be beneficial to exploit this knowledge by enforcing corresponding structure on the activations, that is, to prefer solutions where activations appear as groups matching to a few sources at a time.

Group sparsity allows defining groups of dictionary atoms and constraining the factorisation to use only a small number of groups with active atoms. The technique has been previously employed in some applications, including image classification [7], music separation [8], DNA sequences [9], and automatic speech recognition [10]. In this paper we propose using group sparsity in addition to common L_1 sparsity to produce factorisation solutions, where a narrowed down set of speakers is active at a time. Furthermore, we propose an algorithm which favours the same speakers over the whole duration of an utterance. Sparse activations are shown to produce improved speaker and speech recognition results in a task, where an utterance from an unknown speaker must be recognised among additive noise.

The paper is organised as follows. Section 2 describes the core concepts of spectrogram factorisation and sparse classification. In Section 3 we derive a model and a corresponding iterative update rule to induce consistent group sparsity in utterances comprising multiple observation windows. Experimental set-up on CHiME data is presented in Section 4. Results, discussion and conclusions follow in Sections 5, 6, and 7.

2. Non-negative spectrogram factorisation

Our separation framework is based on representing a mixed observation spectrogram as a linear, non-negative combination of *atoms* — spectrogram segments acquired from sources such as single speakers or background noise. Each atom is modelled with a $B \times T$ magnitude spectrogram matrix, where B is the number of frequency bands and T is *window length* — the number of consecutive frames in an atom. We model noisy speech with J speech and K noise atoms, together forming a *dictionary* (or *basis*) of $L = J + K$ atoms. If we reshape the atoms into length $B \cdot T$ vectors $\mathbf{a}^{\mathrm{s}}_j$ ($j \in [1, J]$) and $\mathbf{a}^{\mathrm{n}}_k$ ($k \in [1, K]$) for speech and noise, respectively, a similarly vectorised observation $\mathbf{y}$ can be estimated as a linear sum

$$\mathbf{y} \approx \sum_{j=1}^{J} \mathbf{a}^{\mathrm{s}}_j x^{\mathrm{s}}_j + \sum_{k=1}^{K} \mathbf{a}^{\mathrm{n}}_k x^{\mathrm{n}}_k \qquad (1)$$

where x^{s}_j and x^{n}_k are the *activation weights* of speech and noise atoms. The same equation can be given in a matrix form as

$$\mathbf{y} \approx \mathbf{A}^{\mathrm{s}}\mathbf{x}^{\mathrm{s}} + \mathbf{A}^{\mathrm{n}}\mathbf{x}^{\mathrm{n}} \qquad (2)$$

where the columns of matrices $\mathbf{A}^{\mathrm{s}}$ and $\mathbf{A}^{\mathrm{n}}$ consist of vectorised speech and noise atoms, and $\mathbf{x}^{\mathrm{s}}$ and $\mathbf{x}^{\mathrm{n}}$ are *activation vectors* for speech and noise, together denoted by vector $\mathbf{x}$ of length L.

In previous work, we have experimented with two different methods to model *observation spectrograms* $\mathbf{Y}$ ($B \times F$), where the number of frames F is larger than T [11]. The first uses $W = F - T + 1$ overlapping windows, each factorised independently. The second, convolutive model is similar but produces a joint spectrogram estimate $\mathbf{\Psi}$ from all window indices simultaneously. Both produce an $L \times W$ *activation matrix* $\mathbf{X}$, each of its columns containing an activation vector for a window index. The previously used cost function to be minimised consists of Kullback-Leibler divergence between the observation spectrogram $\mathbf{Y}$ and its estimate $\mathbf{\Psi}$

$$d_{\mathrm{KL}}(\mathbf{Y}, \mathbf{\Psi}) = \sum_{(y,\psi) \in (\mathbf{Y}, \mathbf{\Psi})} y \log \frac{y}{\psi} - y + \psi \tag{3}$$

and the L_1 norm of $\mathbf{X}$ multiplied elementwise by a sparsity penalty matrix $\mathbf{\Lambda}_1$,

$$f_1 = ||\mathbf{X} \otimes \mathbf{\Lambda}_1||_1. \tag{4}$$

Iterative updates rules to find $\mathbf{X}$ for these costs and for both temporal models can be found in earlier work [3, 11]. In this work, we extend the convolutive model to support group sparsity in addition to basic L_1 sparsity. The same approach for group sparsity also applies to independent window factorisation.

3. Group sparsity for activation matrices

3.1. Multi-column matrix group sparsity

A generalised form of group sparsity can be achieved by using a cost function

$$f_{\mathrm{g}} = ||\sqrt{\mathbf{G}^2 \mathbf{X}^2}||_1 \tag{5}$$

on the activation matrix $\mathbf{X}$. Here $\mathbf{G}$ is a $S \times L$ matrix assigning the L atom indices to S groups with any weights. Square and square root operations are elementwise. The function measures weighted L_2 norms within groups for each window index, produces a $S \times W$ matrix of group 2-norms, and sums them over all groups and window indices. Because in this work we use group sparsity for selection of groups, that is, denoting basic membership without further atom weighting, we simplify the structure by limiting ourselves to assignment matrices of type $\mathbf{G} = \lambda_{\mathrm{g}} \mathbf{G}_{\mathrm{B}}$, where $\mathbf{G}_{\mathrm{B}}$ is a binary matrix denoting atom membership in groups, and λ_{g} is a common weight factor for all chosen atoms. The simplified cost for binary matrices is

$$f_{\mathrm{g}} = \lambda_{\mathrm{g}} ||\sqrt{\mathbf{G}_{\mathrm{B}} \mathbf{X}^2}||_1. \tag{6}$$

However, the given cost function measures group sparsity independently for each window. Although the columns of $\mathbf{X}$ each become sparse on a group level, they may all have different groups active. In our speech recognition task, we expect the same speaker to be active over all window indices within a short observation. Therefore we modify the function to measure the group L_2 norms for summed activity over window indices, $\mathbf{x}_\Sigma = \mathbf{X} \cdot \mathbf{1}$ ($\mathbf{1}$ being an all-one column vector of length W). The cost function becomes

$$f_{\mathrm{g}} = \lambda_{\mathrm{g}} ||\sqrt{\mathbf{G}_{\mathrm{B}} \mathbf{x}_\Sigma^2}||_1. \tag{7}$$

3.2. Combined group and atom sparsity

The equations given in Section 3.1 introduce sparsity over groups, but not over single atoms within a group. Because we have earlier found atom-level sparsity beneficial in SC-based speech recognition as well, both are combined for a cost function that induces sparsity over atoms, yet prefers solutions where the activations come from a sparse set of groups. The total cost function for KL-divergence, group sparsity and L_1 sparsity is

$$f_{\mathrm{tot}} = d_{\mathrm{KL}}(\mathbf{Y}, \mathbf{\Psi}) + \lambda_{\mathrm{g}} ||\sqrt{\mathbf{G}_{\mathrm{B}} \mathbf{x}_\Sigma^2}||_1 + ||\mathbf{X} \otimes \mathbf{\Lambda}_1||_1. \tag{8}$$

3.3. Iterative update algorithm

The total cost function (8) is minimised by initialising all the entries in the activation matrix $\mathbf{X}$ to unity, and then updating it iteratively with an update rule

$$\mathbf{X} \leftarrow \mathbf{X} \otimes \frac{\sum_{t=1}^{T} \mathbf{A}_t^T \overset{\leftarrow(t-1)}{[\frac{\mathbf{Y}}{\mathbf{\Psi}}]}}{\sum_{t=1}^{T} \mathbf{A}_t^T \overset{\leftarrow(t-1)}{\mathbf{1}} + \mathbf{\Lambda}_{\mathrm{g}} + \mathbf{\Lambda}_1}. \tag{9}$$

Here each $\mathbf{A}_t$ is a $B \times L$ matrix containing frame t of all basis atoms. Operator $\leftarrow$ shifts matrix columns left, followed by truncation to W columns. Estimated utterance spectrogram $\mathbf{\Psi}$ is calculated by

$$\mathbf{\Psi} = \sum_{t=1}^{T} \mathbf{A}_t \overset{\rightarrow(t-1)}{\mathbf{X}}. \tag{10}$$

with shifting right ($\rightarrow$) taking place in a $L \times F$ zero-padded matrix. Matrix $\mathbf{\Lambda}_{\mathrm{g}}$ defines the group sparsity cost of each atom and is updated within each iteration based on the activation sum vector. Its columns are identical and are given as

$$\boldsymbol{\lambda}_{\mathrm{g}} = \lambda_{\mathrm{g}} \mathbf{x}_\Sigma \otimes (\mathbf{G}_{\mathrm{B}}^T (\mathbf{G}_{\mathrm{B}} \mathbf{x}_\Sigma^2)^{-1/2}). \tag{11}$$

4. Application to speaker identification and speech recognition

4.1. CHiME data and feature space

To study the potential of group sparsity in finding a sparse combination of sources, we ran experiments on CHiME data, containing GRID command utterances from 34 speakers over family household noises at SNRs ranging from +9 to -6 dB [12]. The utterances follow a linear *verb-colour-preposition-letter-digit-coda* grammar. A default language model utilising 250 sub-word states for the 51 word vocabulary is provided. The data consists of three sets:

- Train: 500 utterances from each speaker without additive noise (‘clean’)
- Development: a set of 600 utterances from all speakers combined, repeated over six SNRs
- Test: as development, but with different utterances and noise content

All audio data, including ‘clean’ sets, has room reverberation. 16 kHz binaural files were used for the experiments. All audio was converted into spectrogram features with $B = 40$ Mel scale spectral bands, 25 ms frame length, 10 ms frame shift, and averaging of the magnitude spectrograms of left and right channels. The bands were linearly scaled using a fixed scaling based on speech training data [3]. Atom length T was set to 25 frames (265 ms).

4.2. Bases and sparsity parameters

We created a 250-atom speech basis for each speaker by modelling the spectrogram context of each state in turn with a $B \times T$ template, based on 300 training utterances per speaker. The procedure is described in earlier work [3, 6]. The concatenated 8500 ($34 \cdot 250$) atom speech basis was used to factorise the remaining 200 training utterances for learning the activation-state mapping matrices needed for sparse classification, in each case with factorisation parameters matching the corresponding test set-up. Mappings were learnt with ordinary least squares regression. During development and test set recognition, a 250-atom noise basis was sampled for each utterance from its noise context and added to the total basis [3].

The binary group sparsity matrix $\mathbf{G}_\text{B}$ ($S \times L$, $S = 34$, $L = 8500 - 8750$) was simply set to 1 for atoms corresponding to speaker s, in other words, for entries 1–250 of group (row) 1, entries 251–500 of group 2 and so forth. The noise atoms at indices 8501–8750, used in all noisy test conditions, did not belong to any group, i.e., the group sparsity constraint was not used for noise. L_1 sparsity weights in matrix $\mathbf{\Lambda}_1$ were kept at 0.1 for entries corresponding to speech and 0.85 for noise as in earlier work. Group sparsity weight λ_g was set to 0.1 based on development set factorisation. All sparsity weights were multiplied by the mean of 1-norms of dictionary atoms to tie the relative weights of KL-divergence and sparsity costs together.

4.3. Recognition experiments

The 3600 test utterances were factorised using the joint 8750-atom basis and 300 iterations of the update rule given in Equation (9). Activation matrices were used for three evaluations:

1. Speaker identification
2. Speech recognition in an external GMM back-end via feature enhancement
3. Speech recognition by sparse classification, that is, determining the state likelihoods from activation weights

All experiments were run with and without the group sparsity penalty, all other parameters remaining identical.

Speaker identification was performed using sparse discriminant analysis (SDA) [6, 13]. Considering the fact that there is only one speaker present in an utterance, we used the summed activity vector $\mathbf{x}_\Sigma$ over an utterance as a feature vector. In order to make the vector invariant to different utterance lengths, the vectors were normalised by the number of windows. The feature vectors from 200 training files per speaker were supplied to an SDA algorithm to find the sparse directions with maximum separability between speakers and minimum variability within speakers. By projecting the 200 vectors from each speaker on sparse discriminant directions, an average model of a speaker was made by simply averaging them. The activity vector of a test segment was also mapped onto SDA directions and dot scoring was employed as the speaker identification score. The number of non-zero elements in SDA was set to 500.

For GMM-based speech recognition, we used the CHiME HTK language model, multi-condition trained GMMs [2], and feature enhancement as in previous work [3]. True speaker identity was exploited in GMM selection in the back-end.

Sparse classification was also performed as in earlier work [3]. Speaker-dependent models were used for Viterbi decoding, although their contribution is limited to transition probabilities, which are highly similar for all speakers.

Table 1: Speaker identification rate (%) comparison for no group sparsity constraint ($\lambda_\text{g} = 0$) and with group sparsity ($\lambda_\text{g} = 0.1$) on the CHiME test set.

SNR	9 dB	6 dB	3 dB	0 dB	-3 dB	-6 dB	avg
$\lambda_\text{g} = 0$	99.8	99.3	98.7	95.5	94.3	82.5	95.0
$\lambda_\text{g} = 0.1$	99.7	99.3	99.2	96.7	93.7	85.7	95.7

5. Results

Results for speaker identification are shown in Table 1. Rates without using group sparsity are shown on the first line, and rates with group sparsity enabled on the second. We observe 0.7% absolute (14% relative) improvement in the average score. Individual SNR scores vary to both directions with debatable significance considering the 600 utterance set size. More on factorisation-based speaker identification results including comparison with GMM baseline can be found in [6].

Table 2 shows the results of speech recognition using factorisation-based enhancement and a GMM back-end. The first two rows contain unenhanced baseline scores for the clean-trained CHiME standard models [12] and multi-condition (MC) trained models [2]. Results for enhancement with different factorisation models are given in the second part of the table. The 8500-atom multi-speaker basis is employed first with L_1 sparsity only, and then with group sparsity enabled. To evaluate the 'oracle' performance obtainable by perfect speaker discrimination, the results on the last row use the true speaker's 250-atom speech basis and the same 250 noise atoms to enhance the signals. We notice that adding group sparsity to multi-speaker basis enhancement produces slight improvements, but only in the noisy end and by a small margin. Neither variant manages to match oracle single-speaker enhancement.

Sparse classification results can be found in Table 3. The same factorisation variants as in enhancement are used for evaluation. This time group sparsity improves the multi-speaker factorisation scores significantly, making them comparable to oracle single-speaker factorisation and classification.

6. Discussion

The results for speaker identification (Table 1) are not entirely conclusive. However, the -6 dB condition is of special interest, because many of its utterances contain loud non-target speech as their background noise. The 18% relative improvement there suggests, that sharpening the distribution of speaker activity manages to remove some interference from non-target speakers. Clean end results are near-perfect to begin with, and there is little confusion between speakers. Consequently no significant changes take place there. Due to the novelty of the approach, further test should be conducted for more conclusive results.

In factorisation-based speech enhancement (Table 2), the speaker identity and state information of atoms is not used in any way — only the spectral features. Therefore features from another speaker are equally valid as long as the spectrograms match, and group sparsity has a limited effect. Improvements in the noisy end can probably be attributed to the non-target background speakers, and the restricted dictionaries' ability to reject secondary identities matching to them. Due to stronger discrimination, such speech is more likely to become modelled with noise atoms as expected. Again, in the clean end differences are limited to only a few test files.

In sparse classification (Table 3), state likelihoods are acquired solely from activation weights and atom labelling. Be-

Table 2: Enhancement-based speech recognition scores (%) over SNRs. Results are shown for clean-trained CHiME baseline models, multi-condition (MC) trained models without enhancement, multi-speaker (MS) enhancement either without or with group sparsity, and finally enhancement by only using the true, single speaker's basis (SS).

SNR	9 dB	6 dB	3 dB	0 dB	-3 dB	-6 dB	avg
GMM baseline scores without enhancement							
CHiME	82.4	75.0	62.9	49.5	35.4	30.3	55.9
MC	91.3	86.8	81.7	72.8	61.1	54.5	74.7
GMM recognition with MC models and enhancement							
MS, $\lambda_g = 0$	92.6	90.3	88.2	84.5	75.6	69.8	83.5
MS, $\lambda_g = 0.1$	92.4	90.4	88.0	85.3	76.2	70.4	83.8
SS	93.0	91.2	90.0	85.2	79.0	72.9	85.2

cause speaker models are trained independently, activations of atoms from other speakers introduce unreliable factors to the final likelihoods. Group sparsity reduces such errors by favouring small sets of active speakers. It is noteworthy that our multi-speaker basis with group sparsity produces recognition rates closely matching informed recognition using the true speaker's basis alone. Because the HMMs can be trained speaker-independently, the whole recognition process becomes speaker-independent over the set of modelled speakers. Together with a robust speaker identification algorithm, the framework provides reliable classification results for both speaker identity and the phonetic content in a scenario, where one unknown speaker from multiple candidates is active at a time.

Concerning the overall rates, it should be noted that the presented framework used small 250-atom speech and noise bases. In other work, we have presented several alternatives for speech and noise modelling [3]. Better results could be achieved by using more accurate speech and noise models, although the efficiency of improved models in conjunction with group sparsity needs to be investigated. While in the presented results speech enhancement was found to perform better than sparse classification, for different bases and features the order may become reversed [3]. Moreover, the two approaches have been found to complement each other in multi-stream recognition [14].

In this study, group sparsity was used for speaker discrimination. However, it is equally feasible to select any sets of atoms for the groups based on their expected co-occurrence. The atom weights in groups need not to be binary either. Different temporal spans can be selected for groups either by choosing an appropriate factorisation spectrogram length, or adjusting the window span used in Equation (7), and then spreading the group sparsity penalty vector (11) accordingly.

7. Conclusions

We proposed using group sparsity in addition to L_1 sparsity in spectral factorisation based noise robust speech recognition in order to limit the number of active speakers from multiple candidates. An iterative update rule was presented for solving convolutive non-negative matrix factorisation with consistent group sparsity over all time indices in an utterance. We found out that the new model manages to narrow down the distribution of speakers, producing marginal but consistent improvements in speaker and speech recognition results. The presented model is generic and allows enforcing also other kinds of group structures in dictionary-based audio spectrogram factorisation.

Table 3: Speech recognition scores (%) with sparse classification. Results are shown for the multi-speaker (MS) basis without and with group sparsity, and then for using the true, single speaker only (SS).

SNR	9 dB	6 dB	3 dB	0 dB	-3 dB	-6 dB	avg
Sparse classification scores							
MS, $\lambda_g = 0$	89.3	87.7	81.5	78.0	68.1	57.9	77.1
MS, $\lambda_g = 0.1$	90.4	88.4	85.7	80.8	73.4	64.3	80.5
SS	89.8	89.0	84.3	81.8	73.9	65.8	80.8

8. Acknowledgements

The work of Rahim Saeidi was funded by the European Community's Seventh Framework Program (FP7/2007-2013) under grant agreement no. 238803.

9. References

[1] P. Smaragdis, "Convolutive Speech Bases and their Application to Supervised Speech Separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 1–14, 2007.

[2] F. Weninger, J. Geiger, M. Wöllmer, B. Schuller, and G. Rigoll, "The Munich 2011 CHiME Challenge Contribution: NMF-BLSTM Speech Enhancement and Recognition for Reverberated Multisource Environments," in *Proc. CHiME workshop*, Florence, Italy, 2011, pp. 24–29.

[3] A. Hurmalainen, J. F. Gemmeke, and T. Virtanen, "Modelling Non-stationary Noise with Spectral Factorisation in Automatic Speech Recognition," *submitted work*, 2012.

[4] S. Rennie, J. Hershey, and P. Olsen, "Single Channel Multi-talker Speech Recognition: Graphical Modeling Approaches," *IEEE Signal Processing Magazine, Special Issue on Graphical Models*, vol. 27, 2010.

[5] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based Sparse Representations for Noise Robust Automatic Speech Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2067–2080, 2011.

[6] R. Saeidi, A. Hurmalainen, T. Virtanen, and D. A. van Leeuwen, "Exemplar-based Sparse Representation and Sparse Discrimination for Noise Robust Speaker Identification," in *Odyssey speaker and language recognition workshop*, Singapore, 2012.

[7] S. Bengio, F. C. N. Pereira, Y. Singer, and D. Strelow, "Group Sparse Coding," in *Proc. NIPS*, 2009, pp. 82–89.

[8] A. Lefèvre, F. Bach, and C. Févotte, "Itakura-Saito Nonnegative Matrix Factorization with Group Sparsity," in *Proc. ICASSP*, Prague, Czech Republic, 2011, pp. 21–24.

[9] L. Meier, S. Van De Geer, and P. Bühlmann, "The Group Lasso for Logistic Regression," *Journal of the Royal Statistical Society: Series B*, vol. 70, pp. 53–71, 2008.

[10] Q. Tan and S. Narayanan, "Novel Variations of Group Sparse Regularization Techniques with Applications to Noise Robust Automatic Speech Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, pp. 1337–1346, 2012.

[11] A. Hurmalainen, J. F. Gemmeke, and T. Virtanen, "Non-negative Matrix Deconvolution in Noise Robust Speech Recognition," in *Proc. ICASSP*, Prague, Czech Republic, 2011, pp. 4588–4591.

[12] H. Christensen, J. Barker, N. Ma, and P. Green, "The CHiME Corpus: a Resource and a Challenge for Computational Hearing in Multisource Environments," in *Proc. INTERSPEECH*, Makuhari, Japan, 2010, pp. 1918–1921.

[13] L. Clemmensen, T. Hastie, D. Witten, and B. Ersboll, "Sparse Discriminant Analysis," *Technometrics*, vol. 54, no. 4, pp. 406–413, 2011.

[14] F. Weninger, M. Wöllmer, J. Geiger, B. Schuller, J. F. Gemmeke, A. Hurmalainen, T. Virtanen, and G. Rigoll, "Non-negative Matrix Factorization for Highly Noise-robust ASR: To Enhance or to Recognize?," in *Proc. ICASSP*, Kyoto, Japan, 2012, pp. 4681–4684.

Publication P6

A. Hurmalainen, J. F. Gemmeke, and T. Virtanen, “Modelling non-stationary noise with spectral factorisation in automatic speech recognition”, in *Computer Speech & Language*, Volume 27, Issue 3, May 2013, pp. 763–779.

Modelling Non-stationary Noise with Spectral Factorisation in Automatic Speech Recognition

Antti Hurmalainen[a,*], Jort F. Gemmeke[b], Tuomas Virtanen[a]

[a]*Department of Signal Processing, Tampere University of Technology, P.O. Box 553, FI-33101, Tampere, Finland*
[b]*KU Leuven, Department ESAT-PSI, Kasteelpark Arenberg 10, 3001 Heverlee, Belgium*

Abstract

Speech recognition systems intended for everyday use must be able to cope with a large variety of noise types and levels, including highly non-stationary multi-source mixtures. This study applies spectral factorisation algorithms and long temporal context for separating speech and noise from mixed signals. To adapt the system to varying environments, noise models are acquired from the context, or learnt from the mixture itself without prior information. We also propose methods for reducing the size of the bases used for speech and noise modelling by 20–40 times for better practical applicability. We evaluate the performance of the methods both as a standalone classifier and as a signal-enhancing front-end for external recognisers. For the CHiME noisy speech corpus containing non-stationary multi-source household noises at signal-to-noise ratios ranging from +9 to -6 dB, we report average keyword recognition rates up to 87.8% using a single-stream sparse classification algorithm.

Keywords: automatic speech recognition, noise robustness, non-stationary noise, non-negative spectral factorization, exemplar-based

1. Introduction

These days we are surrounded by devices and services, which could potentially use speech as their input. Possibly the largest hindrance to widespread adoption of automatic speech recognition (ASR) systems is their limited performance in noisy environments. In everyday situations, the presence of noise can be considered the norm rather than the exception. Therefore robustness against noise is a fundamental requirement for a recogniser intended for common use.

While current state-of-the-art speech recognition systems achieve near-perfect recognition rates on carefully pronounced speech recorded in clean conditions, their performance deteriorates quickly with decreasing signal-to-noise ratio (SNR). Many of the methods proposed for dealing with additive noise focus on increasing the system's sensitivity to the desired patterns over an undefined, roughly uniform noise floor. When the sound level of noise events becomes comparable to that of the target signal, it becomes increasingly important to model noise explicitly. This has been previously accomplished with, for example, model compensation techniques (Acero et al., 2000; Gales and Young, 1996) which allow modelling the interaction of speech and noise. Such techniques have been successfully used to recognise speech in mixtures of multiple speakers, given prior information on each speaker (Hershey et al., 2010).

Since non-negative matrix factorisation (NMF) algorithms were introduced for widespread use (Lee and Seung, 2001), they have been applied to numerous source separation problems. In audio signal processing, NMF has been successfully employed to separate signals consisting of multiple speakers, music, and environmental sounds by modelling a signal as a linear non-negative combination of spectral basis atoms (Heittola et al., 2011; O'Grady and

[*] Corresponding author, tel. +358 41 482 3348, fax +358 3 3115 4989
Email addresses: `antti.hurmalainen@tut.fi` (Antti Hurmalainen), `jgemmeke@amadana.nl` (Jort F. Gemmeke), `tuomas.virtanen@tut.fi` (Tuomas Virtanen)

Pearlmutter, 2007; Schmidth and Olsson, 2006; Smaragdis, 2007; Virtanen, 2007). Given a set of basis atoms (also known as *dictionary*) representing the expected sound sources — in robust ASR, speech and noise — observations can be modelled as a sparse linear combination of atoms. This representation can be used to do speech or feature enhancement, proved useful as a preprocessing step for robust speech recognition (Gemmeke et al., 2011c; Raj et al., 2010; Weninger et al., 2011). Alternatively, when speech atoms are associated with speech classes such as phones, the activations of atoms can provide noise robust likelihoods for hybrid decoding in an approach dubbed *sparse classification* (Gemmeke et al., 2011b; Hurmalainen et al., 2011b).

In the most straightforward approach of spectrograms factorisation, each frame is processed independently. However, in real-world situations, the short-term spectral characteristics of noise events can closely resemble actual speech, making the approach prone to misclassifications. Basic NMF methods have later been extended with prior models (Wilson et al., 2008b), smoothness constraints (Cichocki et al., 2006), temporal dynamic modelling and regularisation (Mysore and Smaragdis, 2011; Wilson et al., 2008a) and adding derivative features to the feature vectors (Van Segbroeck and Van hamme, 2009). Meanwhile, there has been an increasing interest in long context spectro-temporal templates for speech modelling. Example-based methods and longest matching segment searching have been proposed for large vocabulary speech recognition (Sundaram and Bellegarda, 2012; Wachter et al., 2003, 2007), dereverberation (Kinoshita et al., 2011) and denoising (Ming et al., 2011). Multi-frame atoms have also been combined with additive spectral modelling in NMF-based speech separation and enhancement (Smaragdis, 2007; Vipperla et al., 2011; Weninger et al., 2011). In our earlier work, we have found further support for the potential of multi-frame spectrograms as features for robust ASR (Gemmeke et al., 2011b; Hurmalainen et al., 2011b; Hurmalainen and Virtanen, 2012; Virtanen et al., 2010; Weninger et al., 2012). While the benefits of the model have been demonstrated in robust speech recognition, the problem of acquiring effective dictionaries — especially for non-stationary noise — has not been plausibly solved.

In this work, we have three goals. First, we propose a new method for acquiring speech basis atoms from a training set. Thus far, the best recognition accuracy in NMF-based recognition has been obtained by using a large number of atoms, which makes the approach computationally expensive. Therefore methods are needed for selecting smaller sets of atoms that still manage to model speech and noise accurately. The proposed algorithm yields sets of speech basis atoms that are much smaller than the previously employed exemplar sampling methods, which improves the practical applicability of the framework through reduced computational costs.

Second, we propose a method to learn noise basis atoms directly from noisy speech, rather than from pure noise sources. Previous studies show that impressive separation and recognition results can be obtained when accurate prior information on the noises is available. However, when the pre-generated noise model is inaccurate or mismatching, the performance of the methods degrades substantially (Gemmeke et al., 2011b). In our earlier work we employed a technique that samples noise basis atoms from the immediate context of an utterance, similar to the use of a voice activity detector (VAD) to estimate the characteristics of noise during speech inactivity as employed in other noise-robust ASR approaches (Demuynck et al., 2011). Nevertheless, in very noisy conditions a VAD will become unreliable, and for non-stationary noises the estimate acquired during speech inactivity may not match exactly to the noise observed during speech. It is also possible that no reliable source for noise-only segments is available in the first place. In order to overcome these obstacles, we propose to use spectrogram factorisation to learn the noise model using only the noisy speech observation itself as the source for the model. The factorisation algorithm will construct its own noise atoms during separation without prior information or assumptions on the noise events.

The final goal of the paper is to present the current state-of-the-art in spectral factorisation based, single-stream noise robust ASR through the use of spectrogram dynamics and binaural features. Temporal deltas and stereo features are added to the model for increased separation and recognition accuracy.

The rest of the paper is organised as follows: Section 2 describes the spectrogram factorisation tools that are used as the basis for the proposed methods. Section 3 proposes methods for speech and noise model acquisition and adaptation. In Section 4 we present an experimental set-up based on the CHiME noisy speech corpus (Barker et al., 2012) used for public evaluation in CHiME workshop in 2011 (Barker et al., 2011). In Section 5 we present our recognition results, obtained with both sparse classification and front-end speech enhancement based recognition. Discussion and conclusions follow in Sections 6 and 7, respectively.

2. Factorisation-based separation and recognition

2.1. Non-negative spectral modelling

NMF-based separation takes place in a spectro-temporal magnitude domain, where the temporal dimension consists of partially overlapping *frames*, and the spectral dimension of a number of frequency *bands*. In this work, the base unit used for additive modelling is a $B \times T$ spectrogram *window* of B Mel bands and T consecutive frames. These are the dimensions of each observation window in our system, and also of the *atoms*, which form the *basis* for modelling the observation.

We can represent noisy speech as a sum of two parts; a speech model $\hat{\mathbf{s}}$ consisting of speech atoms $\mathbf{a}^{\mathrm{s}}$ weighted by *activations* x^{s},

$$\hat{\mathbf{s}} = \sum_{j=1}^{J} x_j^{\mathrm{s}} \mathbf{a}_j^{\mathrm{s}}, \tag{1}$$

and a noise model $\hat{\mathbf{n}}$ using atoms $\mathbf{a}^{\mathrm{n}}$ and activation weights x^{n},

$$\hat{\mathbf{n}} = \sum_{k=1}^{K} x_k^{\mathrm{n}} \mathbf{a}_k^{\mathrm{n}}. \tag{2}$$

The model uses J atoms for speech and K for noise. The total speech-noise model for noisy observation $\mathbf{y}$ thus becomes

$$\mathbf{y} \approx \hat{\mathbf{s}} + \hat{\mathbf{n}} \tag{3}$$

and the estimated noisy observation

$$\hat{\mathbf{y}} = \sum_{j=1}^{J} x_j^{\mathrm{s}} \mathbf{a}_j^{\mathrm{s}} + \sum_{k=1}^{K} x_k^{\mathrm{n}} \mathbf{a}_k^{\mathrm{n}}, \tag{4}$$

using all in all $L = J + K$ atoms and weight coefficients. For now, we treat basis atoms and the observation as generic feature vectors and ignore their true spectro-temporal ordering, assuming only that they match. All variables are assumed to be strictly non-negative.

The fundamental task is to find the *activation vectors* $\mathbf{x}^{\mathrm{s}}$ (length J) and $\mathbf{x}^{\mathrm{n}}$ (length K), or together simply $\mathbf{x}$, which optimise the model under a chosen quality function. We optimise a cost function consisting of a sum of two factors; first, the generalised Kullback-Leibler (KL) divergence between the observation $\mathbf{y}$ and its approximation $\hat{\mathbf{y}}$

$$d(\mathbf{y}, \hat{\mathbf{y}}) = \sum_{(y_i, \hat{y}_i) \in (\mathbf{y}, \hat{\mathbf{y}})} y_i \log \frac{y_i}{\hat{y}_i} - y_i + \hat{y}_i \tag{5}$$

and second, a penalty term for non-zero activations weighted elementwise by a sparsity vector $\boldsymbol{\lambda}$

$$f(\mathbf{x}) = \|\boldsymbol{\lambda} \otimes \mathbf{x}\|_1 = \sum_{l=1}^{L} \lambda_l x_l. \tag{6}$$

The total cost function to be minimised becomes $d(\mathbf{y}, \hat{\mathbf{y}}) + f(\mathbf{x})$. The first factor measures spectral representation accuracy by generalised KL-divergence, which has been found to perform better than e.g. Euclidean distance or other tested error measures in source separation (Virtanen, 2007). The second factor induces sparsity to the activation vectors, optionally using a customisable weight for each individual basis atom or group of atoms.

2.2. *Sliding window factorisation*

In this work we have used two different approaches for processing utterances longer than the window length T. The first is factorising the utterance in overlapping, independent windows (Gemmeke et al., 2011b). To process an utterance consisting of T_{utt} frames, we advance through it with a step of one frame so that the first window covers frames $[1 \ldots T]$, and the last $[T_{\text{utt}} - T + 1 \ldots T_{\text{utt}}]$. Consequently, we have $W = T_{\text{utt}} - T + 1$ overlapping observation windows over the utterance. Each observation window spectrogram is reshaped to a vector $\mathbf{y}_w$ ($w \in [1, W]$). Similarly, each basis atom is reshaped to a vector $\mathbf{a}_l$ ($l \in [1, L]$). The vectorised observations are collected in a matrix $\mathbf{Y} = [\mathbf{y}_1 \ldots \mathbf{y}_W]$, and the atoms in a *basis matrix* $\mathbf{A} = [\mathbf{a}_1 \ldots \mathbf{a}_L]$. Then we solve for the $L \times W$ *activation matrix* $\mathbf{X}$ so that

$$\mathbf{Y} \approx \mathbf{A}\mathbf{X}, \tag{7}$$

while minimising the cost function defined by equations (5) and (6). This can be achieved by applying iteratively the update rule

$$\mathbf{X} \leftarrow \mathbf{X} \otimes \frac{\mathbf{A}^{\mathrm{T}}(\mathbf{Y}/(\mathbf{A}\mathbf{X}))}{\mathbf{A}^{\mathrm{T}}\mathbf{1} + \mathbf{\Lambda}}, \tag{8}$$

where $\otimes$ is elementiwse multiplication and all divisions are also elementwise. $\mathbf{1}$ is a $\mathbf{Y}$-sized all-ones matrix. $\mathbf{\Lambda}$ is a sparsity penalty matrix defined elementwise for each entry of $\mathbf{X}$, consisting of a λ vector for each observation window.

2.3. *Convolutive factorisation*

An alternative for handling temporal continuity over multi-window observations is *non-negative matrix deconvolution* (NMD), also known as *convolutive non-negative matrix factorisation* (Smaragdis, 2007) or *convolutive sparse coding* (Wang et al., 2011; Wang, 2008). Whereas in the sliding window approach (herefrom called simply 'NMF') each observation window and its corresponding activation vector is an independent entity, in NMD the whole utterance spectrogram $\mathbf{Y}_{\text{utt}}$ is estimated jointly by all activations via convolutive reconstruction. It has been applied earlier to speech separation (O'Grady and Pearlmutter, 2007; Smaragdis, 2007), and to noise-robust speech recognition (Hurmalainen et al., 2011a,b; Vipperla et al., 2011; Weninger et al., 2011).

In this work, we use NMD as in (Hurmalainen et al., 2011a,b). In particular, we only use windows completely within the utterance spectrogram, not ones with their last frames extending beyond T_{utt} as in some implementations. Therefore the activation matrix size is $L \times W$ like in sliding window NMF. The update rule used for activations is

$$\mathbf{X} \leftarrow \mathbf{X} \otimes \frac{\sum_{t=1}^{T} \mathbf{A}_t^T \overset{\leftarrow(t-1)}{[\frac{\mathbf{Y}_{\text{utt}}}{\mathbf{\Psi}_{\text{utt}}}]}}{\sum_{t=1}^{T} \mathbf{A}_t^T \overset{\leftarrow(t-1)}{\mathbf{1}} + \mathbf{\Lambda}}, \tag{9}$$

where each $\mathbf{A}_t$ is a $B \times L$ matrix containing frame t of all basis atoms, and the estimated utterance spectrogram $\mathbf{\Psi}_{\text{utt}}$ is calculated by

$$\mathbf{\Psi}_{\text{utt}} = \sum_{t=1}^{T} \mathbf{A}_t \overset{\rightarrow(t-1)}{\mathbf{X}}. \tag{10}$$

Operators $\overset{\leftarrow i}{(\cdot)}$ and $\overset{\rightarrow i}{(\cdot)}$ denote a matrix shift, where the entries are moved left or right by i columns, respectively.

2.4. *Speech enhancement*

Spectrogram factorisation methods can be used to enhance the input signal before it is passed to a conventional recogniser back-end. Signal enhancement is performed by computing the estimated utterance spectrogram $\mathbf{\Psi}_{\text{utt}}$ as in Equation (10) using the final $\mathbf{X}$ and $\mathbf{A}$ matrices. We also compute an estimated speech spectrogram $\mathbf{\Psi}^{\text{s}}_{\text{utt}}$ by only using the basis atoms and activation rows corresponding to speech. In sliding window NMF the model is similar, except that we average the overlapping window estimates by dividing the frame columns of $\mathbf{\Psi}_{\text{utt}}$ and $\mathbf{\Psi}^{\text{s}}_{\text{utt}}$ by the number of windows contributing to each utterance frame, varying from 1 at the begin and end, to T in the midmost frames.

The clean speech spectrogram estimate is obtained by filtering it in the FFT domain. Because the factorisation model uses Mel-scale spectral resolution, we map the estimates to FFT resolution by inverting the Mel filterbank transform. Denoting the original FFT → Mel scale transform matrix by $\mathbf{M}$, we determine its pseudoinverse $\mathbf{M}^{+}$, and multiply the estimated Mel spectrograms by it from the left. A complex FFT-resolution spectrogram $\tilde{\mathbf{Y}}_{\mathrm{utt}}$ of the original noisy utterance is computed at the temporal resolution of the system. It is then filtered elementwise by the estimated speech/total ratio to get complex speech spectrogram estimate $\tilde{\mathbf{Y}}^{\mathrm{s}}_{\mathrm{utt}}$ as

$$\tilde{\mathbf{Y}}^{\mathrm{s}}_{\mathrm{utt}} = \tilde{\mathbf{Y}}_{\mathrm{utt}} \otimes \frac{\mathbf{M}^{+}\mathbf{\Psi}^{\mathrm{s}}_{\mathrm{utt}}}{\mathbf{M}^{+}\mathbf{\Psi}_{\mathrm{utt}}}. \tag{11}$$

Finally, an enhanced signal is generated with overlap-add synthesis, which inverts the spectrogram derivation.

2.5. *Recognition via sparse classification*

Instead of using factorisation for signal enhancement, the activations can also be used directly for classification (Virtanen et al., 2010). In this approach, dubbed *sparse classification* (SC), speech basis atoms are associated with sequences of speech labels such as HMM-states. The activations of speech basis atoms serve directly as evidence for the associated speech labels, and the combined speech activations yield a state likelihood matrix, which is used in a hybrid HMM-based recogniser. In previous work it was observed that recognition of noisy speech using sparse classification leads to more accurate results than enhancement-based recognition (Gemmeke et al., 2011b). We have also found the performance of SC to improve in some scenarios by replacing the canonical HMM-based labelling of exemplars with atom-state mapping learnt from training set factorisation (Mahkonen et al., 2011).

3. Speech and noise modelling

3.1. *Overview*

To separate sound mixtures, we need atoms to model the contained single source components. In noise robust ASR this means models for pure speech and pure noise. In this section we describe on a general level our methods for generating speech and noise bases from training data, and propose methods for generating noise bases adaptively from the context or from the noisy utterance itself.

3.2. *Pre-sampled exemplar bases*

Both speech and noise bases can be acquired by sampling *exemplars*, instances of spectrograms extracted from the training material as demonstrated in our previous work (Gemmeke et al., 2011b; Hurmalainen et al., 2011b). For speech, this can produce plausible models with high classification capability. For noise, it is not guaranteed that similar sound events will be encountered in actual use cases. In our work on AURORA-2, we saw error rates increasing by up to 60% for mismatched noises (Gemmeke et al., 2011b; Hurmalainen et al., 2011a). Because a noise mismatch degrades the effectiveness of speech-noise separation, and keeping a generic database for all possible noise types would be infeasible, methods for context-sensitive noise modelling are needed for practical applications.

3.3. *Context-based noise sampling*

To reduce the mismatch between observed noise events and the noise basis, we can switch from using a generic noise database to sampling noise exemplars from the nearby context of the utterances to be recognised. It is generally plausible to assume that in ASR the input is continuous, and that there are moments when the target voice is not active. Since exemplars sampled from the immediate noise neighbourhood of utterances are likely to contain sources similar to those in the noisy speech, we exploit these moments without speech activity to update our noise model.

During development of our recognition system, we managed to reduce the error rates by 10–20% by switching from random to context-based noise sampling. The difference depends on the level of mismatch between training data and observed noise. Sampling the local noise context allows more compact bases, lower computational costs, and generally a better match to the noise encountered during speech. The context-based set-up uses annotated ‘oracle’ endpointing to sample its atoms from known noise segments, and exploits both preceding and following temporal context. Although in this work oracle endpointing was used in this work to reduce the number of factors affecting the results and to keep correspondence to earlier work, in (Hurmalainen et al., 2012) preliminary experiments are reported on VAD-based noise segment selection and dynamic basis management for continuous inputs.

3.4. Compact speech bases

Previously we have employed large, semi-randomly sampled speech bases, which typically consist of 4000–5000 exemplars per speaker (Gemmeke et al., 2011b; Hurmalainen et al., 2011b). Experiments have also shown, that further gains in recognition accuracy can be achieved by increasing the number of exemplars. Conversely, a small basis sampled in this manner does not model speech sufficiently well for sparse classification (Gemmeke et al., 2011a). While the large, partially redundant exemplar bases allow accurate modelling of observed speech, they may become difficult to acquire and manage for ASR tasks employing a larger vocabulary.

It is possible to use factorisation algorithms to *learn* the speech bases from training material. This has been previously used for speech separation (Smaragdis, 2007) and speech modelling for denoising (Vipperla et al., 2011; Weninger et al., 2011). Unsupervised learning from diverse speech data will ideally discover recurrent phonetic patterns, which can be used for speech modelling. However, NMF-based algorithms may also separate the spectra of speech patterns into multiple overlapping atoms, or learn short-term events lacking the long temporal context preferred in sparse classification and robust separation. In our preliminary experiments, too much fragmentation has typically taken place in large training set learning for its application to speech basis generation.

To address the issue of basis sizes, in this work we propose modelling speech using *template atoms* with more controlled acquisition and less redundancy. The method is based on constructing an atom for each HMM state in the recognition system, including its typical context. According to HMM state labelling acquired via forced alignment, spectrograms of training data instances corresponding to the chosen state are gathered together, and a characteristic template of the state and its neighbourhood is constructed by averaging. The exact procedure for the CHiME database used in this study is described in Section 4.3.

The variant presented in (Weninger et al., 2011) learns a single basis atom from concatenated instances of one word at a time, making it conceptually similar to the templates used in this work. The main difference lies in our algorithm's capability to model words longer than a single window. By using multiple templates centered around one sub-word state at a time, the system is able to model words of arbitrary length. The partially redundant, state-centered templates can also model speed variations in long word pronunciation by combining multiple activations of sub-word atoms over time.

3.5. Learnt noise bases

Whereas speech training data is generally single-source and can be used as-is to model atomic speech events, noise training data and observations often contain multiple overlapping sources. Therefore learning the noise bases either from noise-only segments or noisy mixtures by applying factorisation algorithms may help us to discover recurrent single-source noise components from mixed signals. In the previously mentioned NMD experiments (Vipperla et al., 2011; Weninger et al., 2011), bases were learnt from segments known to contain only noise. The difference between sampled and learnt atoms primarily depends on the nature of the data. If the co-occurrence of noise sources is low, we can expect the bases to become fairly similar. Some fragmentation of noise events may take place in NMD learning if too many atoms are trained with insufficient sparsity constraints on activation. For strongly multi-source inputs, learning will become more favourable due to its ability to discover atomic sources from mixtures.

A different kind of scenario arises, if no source of pure background noise is available. In this case, we still have an option to learn and separate likely noise artefacts from the noisy utterance itself. Given a sufficiently accurate speech basis, we can factorise a noisy utterance by including self-learning noise atoms in the basis. In this approach, the speech basis is kept fixed, and only the noise part is updated on the fly.

Applying learning to sliding window NMF has some theoretical pitfalls, primarily due to having to learn multiple shifted versions of all noise events. A large learnt basis would be required, which in turn increases the risk of modelling speech with it as well. Preliminary experiments have not produced any promising results on this variant. The NMD model, on the other hand, is well suited for noise learning. Sparsity and a small number of noise atoms act as the restricting factors for isolating new noise events.

Basis learning can be included in the procedure given in Section 2.3. After each iteration of the activation update (9), $\mathbf{\Psi}_{\text{utt}}$ is re-estimated using Equation (10), and the basis is in turn updated by

$$\mathbf{A}_t \leftarrow \mathbf{A}_t \otimes \frac{\frac{\mathbf{Y}_{\text{utt}}}{\mathbf{\Psi}_{\text{utt}}} \overset{\rightarrow(t-1)}{\mathbf{X}}^{T}}{\mathbf{1} \cdot \overset{\rightarrow(t-1)}{\mathbf{X}}^{T}} \quad \forall t \in [1, T]. \tag{12}$$

Learning can be performed for all atoms in the basis or only for a subset of it. In the latter, only the entries of basis and activation matrices corresponding to the atoms to be updated are included in the equation arrays. Afterwards all modified atoms are reweighted to unitary 2-norm.

Ideally, any parts of the spectrogram which cannot be accurately explained with speech exemplars will be captured by the online-learnt noise atoms. This requires some careful calibration to ensure that co-occurring speech features are not captured together with the noise. The primary tool for this is the sparsity weight vector $\boldsymbol{\lambda}$ described in Section 2.1. However, we assume that even cautiously applied noise learning can detect and remove the largest instances of noise, thus filtering out the most harmful artefacts. This is a highly desirable goal for newly encountered noise events, for which we have no prior information.

4. Experimental set-up

4.1. CHiME corpus

For our experiments, we use the CHiME noisy speech database, published in 2010 to address the challenges posed by non-stationary multi-source noise environments (Barker et al., 2012). For its speech content, the database uses the GRID corpus, where 34 different speakers read simple six word command sentences with linear grammar (Cooke et al., 2006). Each utterance follows the syntax *verb-colour-preposition-letter-digit-adverb*. The word classes have cardinality of 4/4/4/25/10/4, respectively. Recognition performance is scored by the percentage of correctly classified *letter* and *digit* keywords. A baseline recogniser employing HTK binaries (Young et al., 2005) with acoustic models trained on clean speech is provided.

The database consists of following sets:

1. Training speech: 500 clean utterances per speaker
2. Training noise: 6+ hours of pure background noise
3. Development set: in total 600 utterances from all speakers, repeated over six SNRs ranging from +9 to -6 dB at 3 dB steps.
4. Test set: As development set, but with different utterances

Test and development utterances are provided in a long noise context as ‘embedded’ files with the utterance locations annotated. Development utterances are also available as clean speech. By ‘clean’ we denote audio without additive noise. All CHiME data is convolved with a room reverberation response, so none of the utterances are truly clean like their original GRID counterparts. All audio is binaural and sampled at 16 kHz.

Additive noise consists of actual household sounds, including appliances, family members, impacts and other sound events. Most of the events are momentary and highly varied, in many cases unique. Different SNRs have been generated by selecting noise segments which produce the desired dB ratio by themselves without scaling. Therefore all SNR-versions of the same development/test utterance contain different noise events.

4.2. Feature space

The feature space used in our experiments consists of magnitude spectrogram segments as described in Section 2.1. The Mel filterbank covers frequencies from 64 to 8000 Hz, divided evenly on a Mel scale with B bands. For the temporal resolution of frames, lengths between 8 and 256 ms have been previously studied (Smaragdis, 2007), and window shift usually varies between 10 and 32 ms. Often a longer frame is used for enhancement than for classification. However, we fix the frame parameters to 25/10 ms for compatibility with CHiME default models and sufficient resolution for sparse classification. In separation and enhancement, it appears that the total duration of atoms, measured in physical time, is more important than temporal resolution within the window (Smaragdis, 2007).

We have previously found repeated evidence for the optimality of window length of 20–30 frames (215–315 ms) for robust enhancement and recognition (Gemmeke et al., 2011b; Hurmalainen et al., 2011a,b). Durations used in other work include 70 ms (Vipperla et al., 2011), 80 ms (Wang, 2008), 176 ms (Smaragdis, 2007), 224 ms (O'Grady and Pearlmutter, 2007) and 256 ms (Weninger et al., 2011, 2012). Based on previous results and a grid search on the development data over a range of T values, we set the NMF window length to 20 frames (215 ms), but use 25 frames (265 ms) for NMD, which appears to favour slightly longer context (Hurmalainen et al., 2011a).

We have achieved improvements by increasing the number of Mel bands from 26 (Hurmalainen et al., 2011b) to 40 (Hurmalainen and Virtanen, 2012). For even larger numbers of frequency bands, the gains were negligible. Therefore B was set to 40 for these experiments.

The factorisation algorithms support processing signals using stereo features by concatenating the features pertaining to each individual channel. In previous work we observed that the use of stereo features only has a minor impact on the separation quality, while it doubles the data size and computational costs (Hurmalainen and Virtanen, 2012). Therefore the results were mostly computed using mono features averaged in the spectral magnitude domain. However, in the same study we found out that augmenting the static features with temporal derivatives ('deltas') similarly as in conventional GMM-based modelling (Young et al., 2005) does improve the recognition rates. Even though the long temporal context of atoms manages to model spectral behaviour over time to some extent by itself, adding explicit delta features will emphasise modulations, which contain significant information on speech and noise events. To generate enhanced signals and recognition results reflecting the current best performance of our framework, stereo features and temporal dynamics as in (Hurmalainen and Virtanen, 2012) were included in the final experiment of this work.

4.3. Basis generation

4.3.1. Speech

In this work, all our speech bases are speaker-dependent, and the knowledge of test speaker identity is exploited by selecting the corresponding speaker's basis. We use two variants; sampling large bases from the training material as described in Section 3.2, and using compact template bases introduced in Section 3.4.

The first method is to sample training utterances semi-randomly (Hurmalainen et al., 2011b). For each speaker, the 500 training utterances are split into 300 for basis generation and 200 for learning the mapping between speech exemplars and speech labels. The utterances selected for basis generation are sampled by extracting windows with a random step of 4–8 frames. The resulting, densely sampled sets of more than 10000 exemplars per speaker are reduced to 5000 while maximising the flatness of included word distribution. This mainly reduces the amount of exemplars from the originally overrepresented non-keyword classes that contain only four word options each. However, no attempt is made to control the exact positioning of exemplars within utterances. They may cover word boundaries, thus modelling specific word transitions.

The second method is based on constructing compact bases of state-centric speech templates. As in the provided CHiME recogniser models, our framework uses 250 speech states (4–10 states per word) to label speech basis atoms. For each state in the system, we select all instances of the word, which contains the chosen state. Based on a forced alignment by the CHiME recogniser, the words are positioned in a length T window with the target state in its midmost frame. We then take the median within each single spectrogram bin over all word instances to generate a prototype of each state and its immediate context. The process is illustrated in Figure 1, where template construction is shown for the third state (out of six) of the word 'green'.

The midmost frames, always representing the nominal state, are most likely to match each other in the spectral domain. Therefore the spectral model is also most consistent in the middle of a template. As the temporal distance increases towards template edges, there is higher variation in the spectrogram content due to differences in pronunciation style, speed and coarticulation. Consequently, the edges fade out when a median is taken over instances. Especially, multiple neighbouring words candidates all have different spectrogram profiles. Consequently the median template model will generally remove the fragments of other words and continuity over word boundaries. For example, in the last training data instance in Figure 1 we can see a high-pitched fricative from a preceding word, whereas very little spectral activity remains in the first frames of the resulting template.

The compact bases cannot model all possible temporal alignments required by independent NMF windows, but they are suited for NMD's temporal model, which can find the best locations for a few temporally sparse activations. By losing word transition modelling and replacing redundant exemplars with median templates, the basis size is reduced to 1/20th of the large NMF bases.

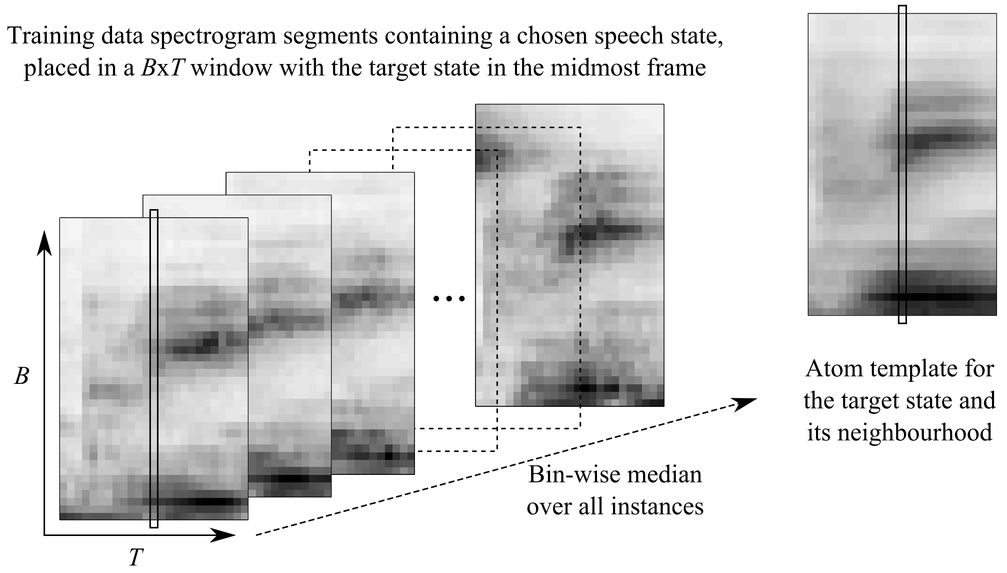

Training data spectrogram segments containing a chosen speech state,
placed in a BxT window with the target state in the midmost frame
B
T
...
Bin-wise median
over all instances
Atom template for
the target state and
its neighbourhood

Figure 1: Forming an atom template for a speech state and its neighbourhood. Training data spectrograms containing the state are placed in a $B \times T$ window, and bin-wise median is taken over the instances. In this example, the third state of word 'green' is modelled with a 40×25 template. In addition to the state itself, a large part of the word is captured as well, thus increasing the temporal context being modelled.

4.3.2. Noise

In this work, we employ three different methods for modelling the additive, non-stationary noise in CHiME data:

1. Context-based sampling of the utterance's noise neighbourhood as presented in Section 3.3 and our earlier CHiME experiments (Hurmalainen et al., 2011b). The 'embedded' wave files are sampled to both directions from the target utterance, and exemplars are extracted at random intervals of 4–7 frames from segments containing only noise. As before, we use 5000 noise exemplars for the NMF experiments. With these parameter settings, approximately 4.5 minutes of noise context got sampled into the basis (from 5–7 minutes of overall audio context with the skipped neighbouring utterances included). The nearest available noise segments were used so the amount of forward and backwards context was roughly symmetric, except at the ends of embedded recording sessions where only one direction is available.
2. The same algorithm, but used to generate a small noise basis of 250 exemplars for NMD. Because less temporal redundancy is required in the NMD model, the sampling interval is increased to 10–15 frames. Still, the overall context covered is reduced to approximately a tenth in terms of physical time span (~ 30 seconds of pure noise data).
3. Finally, we study noise modelling using neither context nor prior knowledge. Instead of passing a pre-generated basis to the factorisation algorithm, we randomly initialise $\lceil T_{\text{utt}}/T \rceil$ noise basis atoms — just enough to cover every frame of an utterance once — and update them in the NMD iteration loop as described in Section 3.5. The on-line updated atoms will adapt themselves to spectrogram patterns not matching to the speech basis, thus learning and modelling noise events found in the mixture.

The generic background training material was not used in any of these experiments. While potentially a sound option in some scenarios, it is debatable if a universal noise basis can be modelled for real world use. For reasons pointed out in Section 3.3, we favour context-aware noise modelling to improve the adaptivity to new noise environments.

4.3.3. Basis weighting

Earlier we have been using two-way normalisation of the basis. Each vectorised atom spectrogram was scaled to unitary Euclidean norm. In addition, the Mel band weights of the full basis were scaled so that the Euclidean norms over all spectral content within each band were equal. To satisfy both conditions together, ten alternating normalisation rounds were performed iteratively for an approximate solution. (Gemmeke et al., 2011b). In this work, we still normalise individual atoms as is preferable for the NMF update rules. However, fixed weights are acquired for Mel bands by gathering all training speech spectrograms, and computing weights which equalise the Euclidean norms over their Mel band content. Using a fixed band weighting profile stabilises and simplifies the model, because the two-way normalisation step can be omitted, and the weighting no longer changes in every noise basis update. When various band weighting methods were compared, the fixed, speech-normalising profile was found to perform comparably to two-way normalisation (Hurmalainen and Virtanen, 2012).

4.4. *Factorisation*

Activation matrices were computed using the update rules described in Section 2. We used CUDA GPU hardware, MATLAB and the GPUmat toolbox (The GP-You Group, 2010) for computation. Single precision variables and 300 iterative updates were used in all experiments.

In many previously reported implementations, the sparsity parameter λ has been set to a fixed value. However, its sparsifying effect is related to the 1-norms of the basis atoms, which will vary as a function of the dimensionality of the feature space. To make the level of sparsity more independent of the window parameters that determine the dimensionality, the penalty weights were set proportionally to the mean of the 1-norms of basis atoms. By conversion from the fixed parameters used in earlier experiments (Hurmalainen et al., 2011b; Hurmalainen and Virtanen, 2012), the sparsity value governing speech basis atoms was set to 0.1 of the mean of norms, and sparsity of noise basis atoms to 0.085. In basis-learning NMD, noise sparsity was increased after brief development data experiments to 0.1 to avoid bias toward the freely adapting atoms and consequently modelling speech with them as well.

4.5. *Decoding*

All our recognition methods are fundamentally based on the CHiME baseline recogniser and its language model. Variants for enhancement and sparse classification are employed as follows.

4.5.1. *Signal enhancement*

In signal enhancement, we synthesise the filtered spectrogram as described in Section 2.4. The enhanced wave files are recognised using HVite and two models with different training. First, we use the default CHiME models trained on reverberated, ‘clean’ training files to produce results compatible with the baseline system. The second system is trained on multi-condition data consisting of the 17 000 clean utterances and the same utterances mixed with random training noise. Mean-only maximum-a-posteriori (MAP) adaptation is used for generating the speaker-dependent models. These models are exactly the same as used in (Weninger et al., 2011) and later in our multi-stream recognition experiments (Weninger et al., 2012).

Neither of these models is retrained on speech data processed with our enhancement framework. Such a task would be laborious, considering that the enhanced output will differ slightly for all factorisation parameters, and that there is no standard training material with noise context as required by our adaptive algorithms. Therefore we only employ generic clean- and multi-condition trained models. A benefit of this choice is that earlier results exist for both models, allowing direct comparison.

In closely integrated recognition systems with matching spectral parameters, it would be possible to use the enhanced Mel scale spectrogram by itself for deriving the MFCC features. However, our separation framework and the two external recognisers all use slightly different parametrisation for their spectral features (e.g. Mel band count and preprocessing filters). Therefore enhanced speech was passed as time domain signals, which are universally accepted by all external recognisers regardless of their internal spectral representation.

4.5.2. *Sparse classification*

For direct classification via speech basis atom weights, we use label matrices representing the probabilities of different speech states over atom duration (Virtanen et al., 2010). In *canonical labelling*, labels are acquired directly from a forced alignment, and the matrices are binary so that for each frame of a speech basis atom only the nominal state is active with weight 1.

However, especially when using speech templates without transition context, some basis atoms may in practise match several different words in the CHiME state model. While phonetically similar, the words are denoted by different states in the system. For example, the first phones of "please" and "place" appear essentially the same. In order to reduce the risk of misclassification due to incorrect or overly strict label associations, we learn the mapping from activations to states by factorising the 200 training utterances not used for the basis, and calculating the mapping matrices using ordinary least squares (OLS) regression (Mahkonen et al., 2011). The non-binary conversion matrices acquired this way are able to model the multiple word associations of some speech atoms, improving the results in scenarios with more phonetic ambiguity (Hurmalainen et al., 2011b).

Preliminary experiments showed that OLS mapping improved the results of small basis experiments, thus this technique was used for the final sparse classification results. For large bases with static features, the results were mixed, with a small overall decrement in average score. With dynamic features included, the results of mapping were uniformly detrimental. Therefore no learnt mapping was used for large basis NMF experiments. The varying benefits of OLS are explained by the accuracy of canonical labels, and the amount of training data. For the large bases with full coarticulation context, the canonical labelling is already reasonably accurate, and no improvements were achieved by learning the mapping from limited training material (200 speaker-dependent utterances). Conversely, the templates constructed from multiple instances have indefinite labels to begin with, and better mapping can be learnt via training factorisation.

The utterances are decoded as described in (Hurmalainen et al., 2011b). In NMF decoding, we normalise the activation vectors of all windows to unitary sum. In NMD's temporal model, the activity levels may vary greatly across windows so no normalisation is applied to the basis activations. The resulting likelihood matrix is passed to a modified CHiME baseline recogniser, which performs the final recognition using the generated likelihoods and the default CHiME language model.

5. Evaluation

5.1. Modelling, factorisation and decoding methods

To compare the different methods for modelling speech and noise, the test set was factorised using three models:

1. Sliding window NMF, 5000 speech and 5000 sampled noise exemplars, $T = 20$ ('Large basis NMF')
2. NMD, 250 speech atoms, 250 sampled noise exemplars, $T = 25$ ('Small basis NMD, sampling')
3. NMD, 250 speech atoms, online-learnt noise model, $T = 25$ ('Small basis NMD, learning')

The 5000-atom sampled speech bases (used in model 1) and 250-atom template bases (models 2 and 3) are described in Section 4.3.1. The three noise models correspond to those described in Section 4.3.2.

Previously, we have got mixed results for applying NMD to large bases (Hurmalainen et al., 2011a,b). For CHiME data, no improvements were seen, while the computational complexity increases significantly. The large bases seem to contain sufficient temporal redundancy for NMF, which in turn produces better results via multiple averaged estimates. Regarding compact bases, the 250+250 atom set-up was tested using both sliding window NMF and NMD. The scores were uniformly worse for NMF than for NMD (0.4–3.5% absolute, 2–20% relative decrement in recognition rates), confirming that the sliding window model is not as well suited for small bases with insufficient temporal alignment variants over the atoms.

All activation matrices acquired from different factorisation types were used for enhancement and recognition with the two GMM-based recognisers; clean-trained original CHiME models ('CHiME') and the multi-condition trained model ('MC'), and also recognised using sparse classification ('SC'). The results are shown in Table 1. The unenhanced baseline performance of the external recognisers is shown on the first rows. Two alternative implementations for NMD enhancement, 'EURECOM' (Vipperla et al., 2011) and 'TUM' (Weninger et al., 2011) are also included on the last rows for comparison.

5.2. Derivative and stereo features

As an additional evaluation, we recomputed the large basis NMF results while including binaural features and temporal dynamics as in (Hurmalainen and Virtanen, 2012). In stereo processing, features were extracted for both channels separately and treated like another set of spectral bands in feature vectors. Temporal dynamics were modelled by applying a delta filter, spanning two frames forward and backwards, to the static magnitude spectrograms. The newly acquired difference spectrogram was split into two parts, one containing positive delta values and another

Table 1: Test set results for different factorisation configurations: large basis NMF, small basis NMD with sampled noise, and small basis NMD with online-learnt noise. All are decoded using feature enhancement (FE) with clean-trained (CHiME) and multi-condition trained (MC) models described in Section 4.5, and sparse classification (SC). Unenhanced baseline scores and two alternative enhancement systems are also shown.

SNR (dB)	9	6	3	0	-3	-6	avg
Baseline scores of FE recognisers (unenhanced)							
CHiME	82.4	75.0	62.9	49.5	35.4	30.3	55.9
MC	91.3	86.8	81.7	72.8	61.1	54.5	74.7
Large basis NMF							
FE, CHiME	92.2	88.8	85.8	80.5	73.3	61.4	80.3
FE, MC	92.8	92.3	90.7	87.6	82.2	75.7	86.9
SC	92.4	90.4	90.0	88.0	79.8	73.8	85.8
Small basis NMD, sampling							
FE, CHiME	91.3	87.0	83.5	76.2	68.2	56.3	77.1
FE, MC	93.0	91.2	90.0	85.2	79.0	72.9	85.2
SC	89.8	89.0	84.3	81.8	73.9	65.8	80.8
Small basis NMD, learning							
FE, CHiME	87.7	83.2	77.2	68.8	60.0	55.4	72.0
FE, MC	91.3	89.8	86.2	80.0	74.2	72.0	82.2
SC	87.8	83.5	79.8	75.0	66.4	60.6	75.5
Alternative NMD enhancement results							
EURECOM[a]	84.6	79.3	69.4	61.8	50.4	43.2	64.8
TUM[b]	90.6	88.3	87.7	84.1	79.2	75.6	84.2

[a] Vipperla et al. (2011)

[b] Weninger et al. (2011)

the absolute values of negative entries in order to keep the features non-negative. In other words, the two derivative spectrograms captured event on- and offsets, respectively. Both were concatenated with the static magnitude features of atoms and observations for separation. However, after acquiring the activation weights, only static magnitudes were used for generating the enhanced spectrogram and signals.

The results for multi-condition trained enhancement (‘FE, MC’) and sparse classification (‘SC’) using extended feature spaces are shown in Table 2. Stereo features and temporal dynamics are first applied each alone and then together. The scores are compared to static-only mono features, and three alternative systems presented in recent literature (Delcroix et al., 2011; Maas et al., 2011; Weninger et al., 2012).

6. Discussion

6.1. Findings

From the results in Table 1, showing the evaluation of different speech and noise modelling methods, we can make the general observation that larger bases and more context information produce better results. This is theoretically sound — the more information available, the better models for individual sources can be constructed. In sparse classification, there is approximately a 5% drop (absolute) in average recognition rate from large basis NMF to small basis NMD, and further to no-prior noise learning. Lower accuracy can already be observed in the cleanest conditions, suggesting that the small bases cannot classify words as accurately as the large bases. However, even the last SC variant performs at least 31%, and on average 44% better than the original CHiME recogniser, measured by relative word error rate reduction.

Interesting results can also be seen in the recognition rate differences between SC and the enhanced signal recognisers. We notice that SC nearly always exceeds the clean-trained CHiME recogniser, while the MC recogniser is

Table 2: Large basis NMF results for static-only mono features, and features with temporal dynamics and stereo channels included. Feature space extensions are applied individually as well as together. Results are shown for multi-condition trained feature enhancement (FE, MC), sparse classification (SC), and three external system combinations reflecting state-of-the-art results on CHiME data.

SNR (dB)	9	6	3	0	-3	-6	avg
Mono, static features only							
FE, MC	92.8	92.3	90.7	87.6	82.2	75.7	86.9
SC	92.4	90.4	90.0	88.0	79.8	73.8	85.8
Stereo, static features only							
FE, MC	93.2	92.2	91.0	87.8	82.4	76.3	87.1
SC	92.4	90.4	90.2	88.4	80.7	73.5	85.9
Mono, static and dynamic features							
FE, MC	93.3	92.1	90.0	87.7	83.1	76.6	87.1
SC	93.0	91.5	90.8	89.2	82.2	76.3	87.2
Stereo, static and dynamic features							
FE, MC	92.9	92.3	90.7	88.2	83.4	77.3	87.5
SC	92.8	91.7	91.1	89.3	83.4	78.6	87.8
Alternative systems for CHiME data							
FAU[a]	95.1	92.6	92.8	88.3	83.3	79.8	88.7
NTT[b]	95.8	94.2	93.7	92.3	88.3	85.6	91.7
TUM/TUT[c]	96.4	95.7	93.9	92.1	88.3	84.8	91.9

[a] Maas et al. (2011)
[b] Delcroix et al. (2011)
[c] Weninger et al. (2012)

mostly better than SC. Especially the small speech basis experiments favour the GMM-based recogniser with robust training. Only 1.7% reduction can be observed in the average score from the 10 000 atom NMF basis to 500 atom NMD. Another 3.0% decrement takes place, when all prior information on noise is removed. Still, enhancement using compact speech modelling and blind noise learning is able to reduce the error rate by up to 38% (relative) in the noisiest end, and by 30% on average in comparison to the same recogniser with unenhanced signals.

The results are also compared to other NMD-based enhancement systems tested on CHiME data. We observe that all our denoising algorithms perform better than the EURECOM approach, where noisy speech was modelled using 100 speech atoms, 100–200 noise atoms from the background data, and 20 atoms from the local context (Vipperla et al., 2011). The results were scored using the standard CHiME recogniser, which therefore should be used as the point of comparison. It is likely that a part of the difference in recognition rates arises from the temporal context, which in our experiments is 20–25 frames (215–265 ms) in comparison to EURECOM's 4 frames (70 ms).

The NMD enhancer used in TUM's CHiME experiments (Weninger et al., 2011) and in our joint work (Weninger et al., 2012) employed 51 speech atoms, 51 noise atoms learnt from the general background, and 256 ms window length. The temporal resolution was 64/16 ms, and the spectral resolution full 1024 FFT bins. The recogniser was the same as the MC model used in this work. We notice that the large basis NMF enhancer performs better than the TUM set-up. Small basis NMD with sampled noise works better in all but the lowest SNRs, and NMD without a noise model only at the highest SNRs. Especially the second case gives some insight to the two systems, which are in many ways similar but also differ in their parametrisation and modelling, primarily in spectral and temporal resolution. It should be inspected further, whether the resolution or the basis generation method plays a larger role in enhancement quality. Differences in the level of sparsity may affect the quality as well.

The final experiment (Table 2) on extended NMF feature spaces reveals more aspects regarding the choice between sparse classification and signal enhancement. Whereas in both static-only set-ups (mono and stereo) features enhancement works better, we notice that including dynamic information improves the SC quality more, making it in turn slightly better. However, the differences are small, so the true order probably depends on implementation details such as external back-end training and the accuracy of atom-to-state mapping in SC. Nevertheless, both recognition methods benefit from dynamic features in separation, especially in the noisy end. The contribution of magnitude-domain stereo information is significantly smaller.

Three alternative recognition systems were also included in Table 2 for comparison. The first one (Maas et al., 2011) is a binaural signal enhancement front-end for a robust Sphinx-4 recogniser employing triphone HMMs. Its noise robustness is generally similar to the proposed system, while its initial clean end recognition rate appears better, probably due to more sophisticated back-end modelling. The NTT approach (Delcroix et al., 2011) combines multiple enhancement and model compensation steps to simultaneously exploit spectro-temporal and spatial information for separation. The TUM/TUT system, also combining multiple streams, consists of GMM recognition, a BLSTM network and a word-spotting version of our sparse classifier (Weninger et al., 2012). This multi-stream system managed to surpass all of its individual streams, and produced the best known average results on CHiME data at the time of writing. We can conclude that system, feature and stream combinations are currently producing state-of-the-art results in noise robust ASR. Factorisation-based methods are well suited for use in such combinations, but other features such as spatial information should also be considered in an efficient overall solution.

6.2. *Computational complexity and costs*

Regarding the computational complexity of factorisation-based speech and noise modelling, we can consider three aspects:

1. Training data requirements
2. Memory allocation
3. Computational costs

We have observed that a large basis of exemplars provides the best accuracy in modelling, and consequently the best recognition results. However, constructing a 5000+5000 atom basis using the approach taken in our NMF experiments requires significant amounts of training data, and for a larger vocabulary the requirements for similar coverage would increase further. Explicit modelling of large word segments and word transitions would require even larger bases, which would only be feasible with dynamic basis management. Fortunately, we have shown that both speech and noise bases can be reduced to a fraction of this size with only a modest decrement in recognition rates. On the other hand, the best results (smallest decrements) were observed using signal enhancement, where some of the training and modelling complexity is shifted on an external back-end.

The memory requirement for NMF bases is $B \cdot T \cdot L$ scalars, which for 10 000 single-precision 40×25 atoms is 40 megabytes. The amount can be reduced significantly by more efficient basis construction, phonetic modelling, and shifting the classification to a conventional recogniser. For example, our 500-atom bases only require 2 megabytes, and with learnt noise atoms even less. Therefore the memory requirements of exemplar-based factorisation are not unbearable for modern devices, including mobile ones.

The computational costs of NMF depend on data sizes, algorithms and naturally the hardware platform. On a dual core E8400 1333 MHz CPU, MATLAB implementation of the large basis (5000+5000 atoms) factorisation takes on average 80.8 seconds per utterance (46× audio duration). On a consumer-grade GeForce GTX260 graphics card, the same computation takes 7.0 seconds per utterance (4.0×). When the basis is reduced to 500 fixed atoms (16× reduction on data size, taking into account the increased window length), NMF execution times become 5.5 seconds (3.1× audio duration) and 0.62 seconds (0.35×) for the described CPU and GPU platforms, respectively. In CPU computing, the speedup factor is close to linear, whereas GPU computing scales better to large arrays due to heavy parallelisation.

Using NMD for factorisation complicates the comparisons. While fixed basis NMF can be computed trivially with elementary matrix operations which also parallelise directly, the NMD speed is highly dependent on algorithm design. The current small basis NMD implementation takes 3–6 seconds per utterance on a GPU, depending on whether basis learning is included. However, the same algorithm for a large basis takes approximately 10 seconds. This highly nonlinear correspondence to problem size illustrates, how the increased computing costs of NMD arise primarily from the overhead of additional algorithm steps. Code optimisation and possibly low-level implementation instead of interpreted MATLAB code would be beneficial in finding out the true performance of NMF and NMD. Nevertheless, it appears ultimately feasible to run the proposed set-ups in real time on parallel platforms.

7. Conclusions and future work

We presented several alternative methods for modelling speech and noise in factorisation-based speech recognition. Local context was used for adaptive noise modelling instead of acquiring a universal noise model from generic training data. The best results were achieved using a large exemplar-based basis consisting of actual instances of training and observation data. Meanwhile, we also demonstrated how significantly smaller bases can be employed for the task with only small losses in quality compared to the reduction factor in model size. Furthermore, we managed to model non-stationary multi-source noise using online-updated atoms without any prior information or context for the noise.

We found additional support for the optimality of 200–250 millisecond window length for both of our recognition methods; signal enhancement for external back-ends and sparse classification based on exemplar labels. When using large bases and dynamic features in addition to static spectra, we achieved better results by sparse classification than by enhancement. However, if the speech bases are reduced to generic templates without word transitions or pronunciation variance, signal enhancement for a multi-condition trained GMM recogniser performed better.

It appears that the current factorisation framework can produce plausible separation results for well-modelled data. Therefore even more effort should be spent on learning compact yet accurate speech and noise models for diverse use cases. The different noise acquisition methods (universal, local context, in-place learning) should be combined to maximise the model accuracy. Preliminary experiments suggest that such combination is indeed feasible, and a noise basis can be updated adaptively in continuous recognition using voice activity detection to locate noise-only segments. Recognition rates comparable to informed noise segment sampling have been achieved by using VAD-based basis adaptation without exploiting any look-forward context (Hurmalainen et al., 2012). For speech, the variations in pronunciation can be possibly handled via clustering or other techniques, which are able to represent the spectro-temporal space volumes with a small number of atoms per phonetic pattern. Switching from word-based to phonetic state models will be eventually needed for large vocabulary recognition.

One important feature type not exploited in this work is the spatial information available in binaural signals. It alone can act as a powerful separation method. Thereby introducing time-domain phase information to the framework might give significant improvements in multichannel recognition.

Regarding final recognition accuracy, there is a lot of potential in multi-stream algorithms, which combine enhancement, sparse classification, and complementary methods (Weninger et al., 2012). Different system combinations should be tested for better joint recognition rates. Especially the clean speech recognition rate, which in our standalone sparse classification is still suboptimal, can be improved by introducing alternative streams to the recogniser. Finally, it would be beneficial to optimise the practical implementation of NMF/NMD algorithms to best exploit current hardware, and thus allow actual deployment of separation-based robust ASR to everyday applications.

8. Acknowledgements

The research of Antti Hurmalainen has been funded by the Academy of Finland. The research of Jort F. Gemmeke was supported by IWT-SBO project ALADIN contract 100049.

References

Acero, A., Deng, L., Kristjansson, T., Zhang, J., 2000. HMM Adaptation using Vector Taylor Series for Noise Speech Recognition, in: Proceedings of the International Conference on Spoken Language Processing (ICSLP), Beijing, China. pp. 869–872.

Barker, J., Vincent, E., Ma, N., Christensen, C., Green, P., 2012. The PASCAL CHiME Speech Separation and Recognition Challenge. Computer Speech and Language (submitted) .

Barker, J., Vincent, E., Ma, N., Christensen, H., Green, P., 2011. Overview of the PASCAL CHiME Speech Separation and Recognition Challenge, in: Proceedings of International Workshop on Machine Listening in Multisource Environments (CHiME), Florence, Italy.

Cichocki, A., Zdunek, R., Amari, S., 2006. New Algorithms for Non-Negative Matrix Factorization in Applications to Blind Source Separation, in: Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Tolouse, France. pp. V–621–624.

Cooke, M., Barker, J., Cunningham, S., Shao, X., 2006. An Audio-visual Corpus for Speech Perception and Automatic Speech Recognition. Journal of the Acoustical Society of America 120, 2421–2424.

Delcroix, M., Kinoshita, K., Nakatani, T., Araki, S., Ogawa, A., Hori, T., Watanabe, S., Fujimoto, M., Yoshioka, T., Oba, T., Kubo, Y., Souden, M., Hahm, S., Nakamura, A., 2011. Speech Recognition in the Presence of Highly Non-stationary Noise Based on Spatial, Spectral and Temporal Speech/Noise Modeling Combined with Dynamic Variance Adaptation, in: Proceedings of International Workshop on Machine Listening in Multisource Environments (CHiME), Florence, Italy. pp. 12–17.

Demuynck, K., Zhang, X., Van Compernolle, D., Van hamme, H., 2011. Feature versus Model Based Noise Robustness, in: Proceedings of INTERSPEECH, Florence, Italy. pp. 721–724.

Gales, M.J.F., Young, S.J., 1996. Robust Continuous Speech Recognition Using Parallel Model Combination. IEEE Transactions on Speech and Audio Processing 4, 352–359.

Gemmeke, J.F., Hurmalainen, A., Virtanen, T., Sun, Y., 2011a. Toward a Practical Implementation of Exemplar-Based Noise Robust ASR, in: Proceedings of European Signal Processing Conference (EUSIPCO), Barcelona, Spain. pp. 1490–1494.

Gemmeke, J.F., Virtanen, T., Hurmalainen, A., 2011b. Exemplar-based Sparse Representations for Noise Robust Automatic Speech Recognition. IEEE Transactions on Audio, Speech, and Language Processing 19, 2067–2080.

Gemmeke, J.F., Virtanen, T., Hurmalainen, A., 2011c. Exemplar-based Speech Enhancement and Its Application to Noise-robust Automatic Speech Recognition, in: Proceedings of International Workshop on Machine Listening in Multisource Environments (CHiME), Florence, Italy. pp. 53–57.

Heittola, T., Mesaros, A., Virtanen, T., Eronen, A., 2011. Sound Event Detection in Multisource Environments Using Source Separation, in: Proceedings of International Workshop on Machine Listening in Multisource Environments (CHiME), Florence, Italy. pp. 36–40.

Hershey, J.R., Rennie, S.J., Olsen, P.A., Kristjansson, T.T., 2010. Super-Human Multi-Talker Speech Recognition: A Graphical Modeling Approach. Computer Speech and Language 24, 45–66.

Hurmalainen, A., Gemmeke, J.F., Virtanen, T., 2011a. Non-negative Matrix Deconvolution in Noise Robust Speech Recognition, in: Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Prague, Czech Republic. pp. 4588–4591.

Hurmalainen, A., Gemmeke, J.F., Virtanen, T., 2012. Detection, Separation and Recognition of Speech From Continuous Signals Using Spectral Factorisation, in: Proceedings of European Signal Processing Conference (EUSIPCO), Bucharest, Romania. pp. 2649–2653.

Hurmalainen, A., Mahkonen, K., Gemmeke, J.F., Virtanen, T., 2011b. Exemplar-based Recognition of Speech in Highly Variable Noise, in: Proceedings of International Workshop on Machine Listening in Multisource Environments (CHiME), Florence, Italy. pp. 1–5.

Hurmalainen, A., Virtanen, T., 2012. Modelling Spectro-Temporal Dynamics in Factorisation-Based Noise-Robust Automatic Speech Recognition, in: Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Kyoto, Japan. pp. 4113–4116.

Kinoshita, K., Souden, M., Delcroix, M., Nakatani, T., 2011. Single Channel Dereverberation Using Example-Based Speech Enhancement with Uncertainty Decoding Technique, in: Proceedings of INTERSPEECH, Florence, Italy. pp. 197–200.

Lee, D.D., Seung, H.S., 2001. Algorithms for Non-negative Matrix Factorization, in: Advances in Neural Information Processing Systems 13, pp. 556–562.

Maas, R., Schwarz, A., Zheng, Y., Reindl, K., Meier, S., Sehr, A., Kellermann, W., 2011. A Two-Channel Acoustic Front-End for Robust Automatic Speech Recognition in Noisy and Rerverberant Environments, in: Proceedings of International Workshop on Machine Listening in Multisource Environments (CHiME), Florence, Italy. pp. 41–46.

Mahkonen, K., Hurmalainen, A., Virtanen, T., Gemmeke, J., 2011. Mapping Sparse Representation to State Likelihoods in Noise-Robust Automatic Speech Recognition, in: Proceedings of INTERSPEECH, Florence, Italy. pp. 465–468.

Ming, J., Srinivasan, R., Crookes, D., 2011. A Corpus-Based Approach to Speech Enhancement From Nonstationary Noise. IEEE Transactions on Audio, Speech, and Language Processing 19, 822–836.

Mysore, G.J., Smaragdis, P., 2011. A Non-negative Approach to Semi-Supervised Separation of Speech from Noise with the Use of Temporal Dynamics, in: Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Prague, Czech Republic. pp. 17–20.

O'Grady, P.D., Pearlmutter, B.A., 2007. Discovering Convolutive Speech Phones using Sparseness and Non-Negativity Constraints, in: Proceedings of ICA, London, UK. pp. 520–527.

Raj, B., Virtanen, T., Chaudhure, S., Singh, R., 2010. Non-negative Matrix Factorization Based Compensation of Music for Automatic Speech Recognition, in: Proceedings of INTERSPEECH, Makuhari, Japan. pp. 717–720.

Schmidth, M.N., Olsson, R.K., 2006. Single-channel Speech Separation using Sparse Non-negative Matrix Factorization, in: Proceedings of the International Conference on Spoken Language Processing (ICSLP), Pittsburgh, Pennsylvania, USA. pp. 2614–2617.

Smaragdis, P., 2007. Convolutive Speech Bases and their Application to Supervised Speech Separation. IEEE Transactions on Audio, Speech, and Language Processing 15, 1–14.

Sundaram, S., Bellegarda, J., 2012. Latent Perceptual Mapping with Data-Driven Variable-Length Acoustic Units for Template-Based Speech Recognition, in: Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Kyoto, Japan. pp. 4125–4128.

The GP-You Group, 2010. GPUmat User Guide, version 0.27.

Van Segbroeck, M., Van hamme, H., 2009. Unsupervised Learning of Time-Frequency Patches as a Noise-robust Representation of Speech. Speech Communication 51, 1124–1138.

Vipperla, R., Bozonnet, S., Wang, D., Evans, N., 2011. Robust Speech Recognition in Multi-Source Noise Environments using Convolutive Non-Negative Matrix Factorization, in: Proceedings of International Workshop on Machine Listening in Multisource Environments (CHiME), Florence, Italy. pp. 74–79.

Virtanen, T., 2007. Monaural Sound Source Separation by Non-Negative Matrix Factorization with Temporal Continuity and Sparseness Criteria. IEEE Transactions on Audio, Speech, and Language Processing 15, 1066–1074.

Virtanen, T., Gemmeke, J., Hurmalainen, A., 2010. State-based Labelling for a Sparse Representation of Speech and Its Application to Robust Speech Recognition, in: Proceedings of INTERSPEECH, Makuhari, Japan. pp. 893–896.

Wachter, M.D., Demuynck, K., Compernolle, D.V., Wambacq, P., 2003. Data-Driven Example Based Continuous Speech Recognition, in: Proceedings of EUROSPEECH, Geneva, Switzerland. pp. 1133–1136.

Wachter, M.D., Matton, M., Demuynck, K., Wambacq, P., Cools, R., Compernolle, D.V., 2007. Template-based Continuous Speech Recognition. IEEE Transactions on Audio, Speech, and Language Processing 15, 1377–1390.

Wang, D., Vipperla, R., Evans, N., 2011. Online Pattern Learning for Non-Negative Convolutive Sparse Coding, in: Proceedings of INTERSPEECH, Florence, Italy. pp. 65–68.

Wang, W., 2008. Convolutive Non-Negative Sparse Coding, in: Proceedings of IJCNN, Hong Kong. pp. 3681–3684.

Weninger, F., Geiger, J., Wöllmer, M., Schuller, B., Rigoll, G., 2011. The Munich 2011 CHiME Challenge Contribution: NMF-BLSTM Speech Enhancement and Recognition for Reverberated Multisource Environments, in: Proceedings of International Workshop on Machine Listening in Multisource Environments (CHiME), Florence, Italy. pp. 24–29.

Weninger, F., Wöllmer, M., Geiger, J., Schuller, B., Gemmeke, J.F., Hurmalainen, A., Virtanen, T., Rigoll, G., 2012. Non-negative Matrix Factorization for Highly Noise-robust ASR: To Enhance or to Recognize?, in: Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Kyoto, Japan. pp. 4681–4684.

Wilson, K.W., Raj, B., Smaragdis, P., 2008a. Regularized Non-Negative Matrix Factorization with Temporal Dependencies for Speech Denoising, in: Proceedings of INTERSPEECH, Brisbane, Australia. pp. 411–414.

Wilson, K.W., Raj, B., Smaragdis, P., Divakaran, A., 2008b. Speech Denoising Using Nonnegative Matrix Factorization with Priors, in: Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Las Vegas, Nevada, USA. pp. 4029–4032.

Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., Woodland, P., 2005. The HTK Book Version 3.3. Cambridge University Press.

Publication P7

A. Hurmalainen and T. Virtanen, “Acquiring Variable Length Speech Bases for Factorisation-Based Noise Robust Speech Recognition”, in *Proceedings of the 21st European Signal Processing Conference (EUSIPCO)*, Marrakech, Morocco, 9.–13. September 2013, pp. 1495–1499.

ACQUIRING VARIABLE LENGTH SPEECH BASES FOR FACTORISATION-BASED NOISE ROBUST SPEECH RECOGNITION

Antti Hurmalainen, Tuomas Virtanen

Tampere University of Technology, P.O. Box 553, 33101 Tampere, Finland

ABSTRACT

Studies from multiple disciplines show that spectro-temporal units of natural languages and human speech perception are longer than short-time frames commonly employed in automatic speech recognition. Extended temporal context is also beneficial for separation of concurrent sound sources such as speech and noise. However, the length of patterns in speech varies greatly, making it difficult to model with fixed-length units. We propose methods for acquiring variable length speech atom bases for accurate yet compact representation of speech with a large temporal context. Bases are generated from spectral features, from assigned state labels, and as a combination of both. Results for factorisation-based speech recognition in noisy conditions show equal or better separation and recognition quality in comparison to fixed length units, while model sizes are reduced by up to 40%.

Index Terms— Spectral factorization, speech recognition, noise robustness

1. INTRODUCTION

Speech contains phonetic units of varying lengths, ranging from single phones to their combinations, syllables, words and complete phrases. Statistical analysis of speech reveals correlation in its temporal behaviour spanning hundreds of milliseconds, decreasing gradually with no strict upper limit [1]. Meanwhile, physiological studies and listening tests have shown that temporal modulations at under 12 Hz (period of 83 ms or more) are crucial for speech intelligibility [2].

Conventional automatic speech recognition (ASR) systems typically use frames of approximately 25 ms as their features, and Markovian state transition models which only consider temporal context of one frame. The approach is computationally efficient and sufficient for single phone classification, but fails to model the long term temporal behaviour motivated by natural speech structures and human hearing. Especially in noisy conditions short-term spectra become unreliable as features for classification. Separating and recognising sources from a single frame is often an ill-posed problem. While partial alleviation can be achieved by including delta and acceleration features to frame spectra, the context still remains limited, and extended temporal connectivity actually violates the Markovian model assumption [1]. Due to these limitations and the need for more robust models, there is increasing interest towards long context spectrogram modelling in ASR [3].

Several approaches have been proposed for increasing the context of speech models. TRAPs features observe long term temporal behaviour of a few spectral bands [4]. HAC models quantise frame level audio events into classes and form histogram vectors summarising the events in variable length words [5]. Phonetic segmentation of speech has been discussed and demonstrated in literature [6], although evaluation has usually consisted of comparison to manual segmentation with no application to ASR. Longest segment matching has been applied to dereverberation [7] and robust ASR [8]. Increased context has also been used in deep belief networks with optimal results gained at contexts of 110–270 ms [9].

Using spectro-temporal atoms spanning 200–300 ms has been shown to provide high separation quality and noise robustness with methods based on *non-negative matrix factorisation* (NMF) [10, 11, 12]. However, the exact choice of window length has proven difficult. Increasing the context will improve robustness. On the other hand, it increases the complexity of modelled spectro-temporal patterns, thus requiring more atoms for the same data. Furthermore, fixed atom length does not correspond to the large variation of acoustic units occurring in real world speech and noise.

While virtually all studies on NMF thus far have concentrated on fixed atom length models, more recently variable length modelling has also been proposed. Yılmaz et al. used combination of factorisation passes with multiple fixed length dictionaries [13]. Although a promising step towards variable length modelling, using multiple large dictionaries may prove impractical. Meanwhile, Wang and Tejedor have proposed a model for employing different atom lengths simultaneously in convolutive NMF (also known as NMD) [14], and presented an introductory experiment on two-speaker separation.

In this work we extend variable length NMD modelling to robust ASR, and propose methods for acquiring compact speech bases with a preference for long context, yet able to model units of any length. We employ two data sources for finding units; unannotated spectral features, and state labels acquired from a language model via forced alignment. The two sources are also used in conjugation. Models are evaluated in a noisy ASR task using the 1st CHiME Challenge corpus [15]. Factorisation-based representation is used for

T. Virtanen has been funded by the Academy of Finland, grant #258708.

feature enhancement for an external back-end, and for ASR directly from atom activations. First we introduce the fundamentals of spectral factorisation. The proposed acquisition method is presented in Section 3. In Section 4 we describe the evaluation set-up and experiments. Results are listed and discussed in Section 5, whereafter we conclude in Section 6.

2. SPECTROGRAM FACTORISATION

In spectrogram factorisation, the goal is to model a mixed *observation spectrogram* $\mathbf{Y}$ as a sum of separated source spectrogram estimates, which in robust ASR comprise speech $\mathbf{\Psi}^{\mathrm{s}}$ and noise $\mathbf{\Psi}^{\mathrm{n}}$. The dimensions of each utterance spectrogram are $B \times T_{\text{utt}}$, where B is the number of spectral bands and T_{utt} is the number of frames. The estimates are constructed by weighted summing of *atom spectrograms* $\mathbf{A}_l$ ($B \times T_l$). Atoms are indexed by l from 1 to *basis size* L. Whereas in earlier work the *atom length* T_l has been a constant [11, 12], in this work we allow it to vary between atoms.

Each utterance spectrogram estimate $\mathbf{\Psi}$ is a convolutive sum of atom spectrograms, weighted by a $L_G \times W$ *activation matrix* $\mathbf{X}$. L_G is the number of atoms belonging to set G of the source(s) being modelled. W is the number of permitted *window indices*, equal or less than T_{utt}. The convolutive reconstruction formula for a spectrogram $\mathbf{\Psi}_G$ is

$$\mathbf{\Psi}_G = \sum_{l \in G} \sum_{t=1}^{T_l} \mathbf{A}_{l,t} \overset{\rightarrow(t-1)}{\mathbf{X}_l}, \tag{1}$$

where $\mathbf{A}_{l,t}$ denotes the t^{th} frame column of atom l, $\mathbf{X}_l$ is the l^{th} row vector of $\mathbf{X}$, and operator $\rightarrow$ shifts it right by $t-1$ columns in a length T_{utt} zero-padded array to make all partial matrices to be summed $B \times T_{\text{utt}}$. The method is otherwise similar to commonly used convolutive modelling [16], except that the atom length T_l can be given separately for each atom.

Assuming a pre-generated supervised basis $\mathbf{A}$, the factorisation task consists of finding the activation matrix $\mathbf{X}$ for a chosen quality function. After solving $\mathbf{X}$, it is used either for estimating source spectrograms as above, or directly as a classifier by observing the activated atoms and their supplementary label information. Both methods have been used for robust ASR. Their details can be found in earlier work [10, 12], and are also given briefly in the following sections.

3. COMPACT VARIABLE LENGTH BASES

In previous work, fixed length atoms have been acquired by sampling randomly a large amount of *exemplars* [10], or by constructing templates for each word of a small vocabulary [11, 12]. However, both methods may prove problematic when real world speech must be modelled. In order to cover spectro-temporal patterns of speech with a compact set of atoms, we propose an algorithm which aims at discovering recurring events of variable length with a preference for long units. The algorithm is based on searching for *clusters* of speech segments which match each other.

First, let us define a similarity measure c between two frame feature vectors $\mathbf{f}^{(i)}$, $\mathbf{f}^{(j)}$. The frames are considered *matching* if their c value exceeds a given threshold θ. Similarly, two *sequences* of length N, $[\mathbf{f}_1^{(i)} \ldots \mathbf{f}_N^{(i)}]$ and $[\mathbf{f}_1^{(j)} \ldots \mathbf{f}_N^{(j)}]$ are considered matching if all their mutual vector pairs $\mathbf{f}_n^{(i)}$, $\mathbf{f}_n^{(j)}$ match. Because the atoms in NMD are rigid with no time warping, it is crucial that sequences match throughout their duration.

We consider two different data sources for finding matches. First, we observe the spectral features of frames, denoted by $\mathbf{s}$. Spectral matching can be defined by any similarity measure, but in this work we use straightforward dot product

$$c_{\mathrm{s}}(i,j) = \mathbf{s}^{(i)} \cdot \mathbf{s}^{(j)} \tag{2}$$

between L_2-normalised spectrum vectors. The largest possible spectral similarity is thus 1.

The second method is using phonetic state labels acquired from word transcriptions with forced alignment. Each frame in training data is given a label denoting its membership in exactly one language model state q of total Q states. We define the similarity of states in frames i and j as

$$c_l(i,j) = \begin{cases} \gamma_{\text{full}} & \text{if } q^{(i)} = q^{(j)} \\ \gamma_{\text{part}} & \text{if } |q^{(i)} - q^{(j)}| = 1 \\ \gamma_{\text{none}} & \text{otherwise} \end{cases} \tag{3}$$

The midmost ‘partial match’ is true if the states follow each other in a linear language model, thus allowing minor errors in alignment. Finally, the similarity measures may be combined by using a merging function. In this work the function is the sum of coefficients,

$$c_m(i,j) = c_{\mathrm{s}}(i,j) + c_l(i,j) \tag{4}$$

The relative significance of spectral and state similarity can be defined via γs and the threshold θ.

Clustering is implemented with a greedy longest-first search. Starting from the largest allowed atom length T_{max}, we find pairwise matching sequences from training data. If a sequence is found with a sufficient number of matches to other sequences, these instances form a cluster and further an atom. The contained frame ranges are flagged as taken. Then the algorithm continues clustering, reducing the atom length T when clusters of chosen size can no longer be found. Halting can be defined e.g. by the number of extracted atoms, percentage of modelled training data, or a minimum atom length T_{min}. Although the greedy algorithm does not guarantee global maximisation of atom lengths, it is practically viable and produces a basis of recurring spectral patterns in a descending order of length and frequency of occurrence.

4. EXPERIMENTAL SET-UP

4.1. Data set and features

For the experiments, we used the GRID-based 1st CHiME Challenge corpus [15]. Its speech consists of six-word command utterances following a linear *verb-colour-preposition-letter-digit-adverb* grammar. Word classes have cardinalities 4/4/4/25/10/4 respectively, totalling to 51 words. The task is to recognise 'letter' and 'digit' keywords. There are 34 speakers, and a 500-utterance training set is provided for each. Speaker identity is assumed known in recognition. Noisy development and test sets both contain 600 utterances mixed with highly non-stationary room noise at six SNRs ranging from +9 to -6 dB. All audio data contains room reverberation.

All binaural source audio was converted into 40-band mel-spectral features using 25 ms frames with 10 ms shift, and averaged into mono. Spectral bands were equalised with fixed band weights derived from training data [12]. Default CHiME language model comprising 250 sub-word states and its forced alignment were used to assign state labels to frames.

4.2. Frame correlation functions

Three correlation variants were used for clustering speech frames in basis acquisition:

1. Spectral features only ('spect')
2. State labels only ('label')
3. Combination of the two ('comb')

The spectral space employed mel magnitudes with square root compression and augmented delta features derived from a five-frame window [12]. Spectral similarity c_{s} was measured as the dot product of 2-normalised vectors. The features were chosen for invariance to absolute loudness, while retaining the temporal dynamics of speech. In purely spectral acqusition θ was set to 0.89 and state correlation c_l was 0.

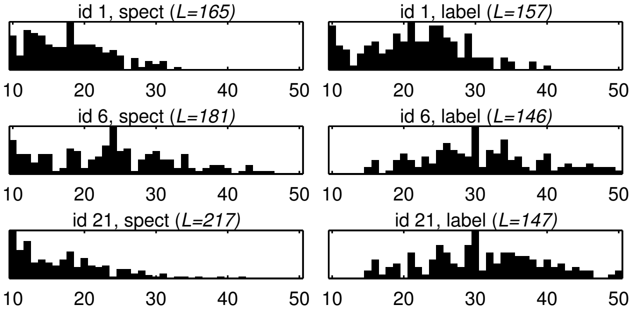
In solely label-based acquisition, c_{s} was in turn set to 0. Label correlation values were $\gamma_{\mathrm{full}} = 2$, $\gamma_{\mathrm{part}} = 1$ and $\gamma_{\mathrm{none}} = 0$. Threshold θ was set to 1 with an additional constraint that the mean correlation between sequences was over 1.8. In other words, all state pairs must correlate at least partially, and 80% of them must match perfectly. As the algorithm has no access to spectra, we require relatively strict state sequence similarity with a small allowance for fluctuations.

Combined acquisition used the same spectral correlation with θ increased to 0.92. However, c_l used $\gamma_{\mathrm{full}} = 0.06$, $\gamma_{\mathrm{part}} = 0.03$ and $\gamma_{\mathrm{none}} = 0$. Parameters were tuned in 0.01 steps using development data.

4.3. Basis acquisition

After defining the frame correlation functions, speech bases were acquired from training data for each speaker separately as follows. Starting from T_{max}, all pairwise matching sequences were searched from training data. Because in GRID data each word is chosen randomly from its class and no word transition is more likely than another, we restricted learning to clusters modelling a single word each. A cluster was selected if it contained at least 25% of the modelled word's instances. At each window length, all such clusters were extracted in a descending order of relative size, whereafter T was reduced by one frame. The process was halted either by reaching the minimum length T_{min}, or if at least 75% of non-silent training frames were already covered. Sequences were allowed to span over silent frames (defined by spectral energy) to model e.g. stop consonants, but not to end in one, as such cases could be modelled with a shorter atom instead.

Fig. 1. Histograms of atom lengths in selected speakers' bases for spectrum-based (left) and label-based (right) acquisition. L is the total number of atoms.



Each cluster was converted into a speech atom by averaging its mel magnitude spectrograms binwise. In addition, the preceding and succeeding 2 frames were included in atoms, because their magnitude content is implied by delta features. Original T_{max} and T_{min} were set to 46 and 6, thus final atom lengths ranged from 50 to 10 frames. A few examples of atom length histograms within individual speakers' bases are shown in Figure 1. For now, we can notice that the whole range is employed in different variants, and the distribution depends heavily on the speaker and the method. Further analysis is given later in Section 5.

A summary of the generated bases is shown in Table 1. For each generation method; spectrum-based ('spect'), label-based ('label') and combined ('comb'), we list the statistics of atom counts, total frame counts, and average atom lengths of the 34 speaker-dependent bases. Previously used fixed-length bases ('fixed') with exactly 250 length 25 atoms per speaker are included for comparison [12].

4.4. Factorisation and recognition

The factorisation and recognition framework mostly follows small basis experiments described in [12]. A joint speech+noise basis was formed from a variable number of speaker-dependent speech atoms (see Table 1), and 250 noise atoms sampled from the context of test utterances. Activation matrices $\mathbf{X}$ were solved with variable-length NMD described in Section 2 [14]. 300 iterations were used as before, and an L_1 sparsity penalty was applied for better separation and classification. Where possible, parameters were set as in the 250+250 fixed length atom experiments in [12]. Especially the fixed length noise atoms were replicated exactly to study the contribution of new speech models alone.

For decoding and recognition, we used two methods. The first is sparse classification (SC) via activation weights and

Table 1. Statistics of the 34 speaker-dependent speech bases, listed for all acquisition methods. Number of atoms in a basis, amount of contained frames, and average atom length are reported as minimum, mean and maximum values over speakers. The reference method always uses 250 length 25 atoms.

method	atom count			frame count			avg atom length		
	min	mean	max	min	mean	max	min	mean	max
spect	160	190	237	2941	3793	4659	17.0	20.0	23.6
label	135	150	181	3346	4167	5052	21.6	27.9	33.9
comb	157	182	232	3151	4027	4903	18.6	22.2	25.8
fixed	250			6250			25		

Table 2. Keyword recognition rates (%) and SDRs (dB) for unenhanced signals, proposed, and reference basis acquisition methods. Results are averages over noisy conditions from +9 to -6 dB. The best result among small basis methods (spect, label, comb, fixed) for each set is highlighted.

method	development set			test set		
	SC	FE	SDR	SC	FE	SDR
unenh	-	74.6%	-0.72 dB	-	74.7%	-0.78 dB
spect	79.4%	85.1%	7.87 dB	79.9%	85.4%	8.50 dB
label	78.3%	**85.6%**	**8.80 dB**	78.9%	**85.6%**	**8.86 dB**
comb	**79.7%**	85.3%	8.54 dB	80.3%	85.5%	8.58 dB
fixed	78.0%	84.8%	8.57 dB	**80.8%**	85.2%	8.62 dB
large	85.9%	86.7%	9.49 dB	85.8%	86.8%	9.55 dB

atom labels [10, 12]. In this method, a $Q \times T_{\text{utt}}$ state likelihood matrix is generated similarly to the spectrogram estimate of Equation (1) using $Q \times T_l$ label matrices assigned to speech atoms. Labels were learnt by partial training set factorisation and ordinary least squares regression between activations and utterance state content [12]. Final likelihood matrices were decoded directly using the default CHiME HMMs.

The second method is feature enhancement (FE) by using the ratio $\boldsymbol{\Psi}^{\text{s}}/(\boldsymbol{\Psi}^{\text{s}} + \boldsymbol{\Psi}^{\text{n}})$ of speech-only and total spectral reconstructions from Equation (1) as a time-varying filter for the original utterance spectrogram [12]. The enhanced signal was passed to a multi-condition trained robust GMM backend, previously used in [11, 12]. Details of both methods can be found in earlier work [12].

5. RESULTS AND DISCUSSION

Speech recognition and enhancement results for each modelling method are listed in Table 2 as keyword recognition rates for sparse classification (SC) and feature enhancement (FE), and signal-to-distortion ratio (SDR) of enhanced utterances measured with the BSS Eval toolkit [17]. Shown values are averages over noisy conditions and given for development and test sets separately. The first line contains baseline results for unenhanced signals. The next three lines correspond to similarity measures defined in Section 4.2 for variable length modelling. Results for previous 250+250 atom fixed-length modelling ('fixed'), and significantly larger 5000+5000 atom NMF bases ('large') are also included for comparison [12].

First, we can observe from Table 1 that in each measure of basis sizes, approximately 10–25% deviations take place between speakers from the mean to minimum and maximum values, illustrating the model's adaptivity. Mean atom count is reduced by 24.0–40.0% and mean frame count by 33.3–39.3% in comparison to fixed-length bases. Mean atom lengths vary significantly between speakers and methods. Spectral models produces more and shorter atoms than labels. Source combination generally falls inbetween.

Although the statistics ultimately depend on the similarity functions and clustering parameters, the observed trend can be justified by properties of the functions. Feature-only modelling will discover recurring spectral units, which are often shorter than whole words due to coarticulation and natural variation in pronunciation. State-only models are based on forced alignment, which always produces a similar sequence regardless of phonetic variation. It only observes variations in pacing, which are more consistent for any given speaker. This can be seen in the atom length histograms of Figure 1. Speaker 1 is fast and produces short atoms for both methods. Speaker 6 is slow and clear, hence both bases have longer atoms. Speaker 21 is relatively slow but very melodic. In this case, the feature-based atoms are shortest in the whole set, whereas state-based atoms are among the longest.

Regarding the separation and recognition results of Table 2, there is some variation between methods for different result metrics. While separation measured by SDR is either above or below the previous 'fixed' method, FE-based ASR results improve uniformly. Gains are small, but it should be noted that the gap to 20 times larger exemplar models ('large') is only $< 2\%$. The performance of SC is harder to analyse, because the results for development and test set differ greatly for the fixed-length model, while the proposed methods are more consistent. One contributing factor is that SC for CHiME data depends heavily on keyword modelling. In the previous model, at least four atoms per word were guaranteed, whereas the proposed method has no such constraints. In individual SNR level scores (not shown), the proposed methods had slightly lower clean end classification quality but higher robustness towards low SNRs. Separation and classification also have partially conflicting goals with the former preferring long atoms, but the latter requiring also short atoms which bear a higher risk of confusion with noise.

The main benefit of the presented method is that it can adapt to any vocabulary and speaking style, unlike the previous model which assumed long context implied by sub-word labels of small vocabulary and required defining the window length explicitly. Although a small vocabulary task was used here for simplicity of presentation and easier comparison to earlier work, we have already employed the methods — both feature- and state-based — successfully to compact modelling of medium vocabulary speech [18]. Regarding complexity, the basis acquisition time for this task was < 30 minutes per speaker using MATLAB code and an E8400 dual-core desktop PC. For larger corpora, computation of full similarity may become slow, thus pre-classification and approximate methods may become recommendable.

While in this work a fixed-length noise model was used to limit the number of parameter changes, variable-length methods are equally applicable to noise, where the variation between unit lengths may be even greater than for speech.

6. CONCLUSIONS

We proposed methods for acquiring variable-length long-context speech bases for noise robust speech separation and recognition. Spectral features, state labels, and a combination of both were used for clustering speech patterns to atoms via longest-first segment search. Applied to 1st CHiME Challenge data, the methods produced speaker-adaptive bases with atom lengths ranging from 10 to 50 frames. We managed to reduce model sizes by up to 40% from already compact fixed-length bases, while achieving similar or better separation and speech recognition results. The presented methods can be used to model large vocabulary speech and non-stationary noise for better applicability to real world ASR scenarios.

7. REFERENCES

[1] O. Räsänen and U.K. Laine, "A method for noise-robust context-aware pattern discovery and recognition from categorical sequences," *Pattern Recognition*, vol. 45, no. 1, pp. 606–616, 2012.

[2] T.M. Elliott and F.E. Frédéric, "The Modulation Transfer Function for Speech Intelligibility," *PLoS Computational Biololgy*, vol. 5, no. 3, pp. e1000302, 2009.

[3] T.N. Sainath, B. Ramabhadran, D. Nahamoo, D. Kanevsky, D. Van Compernolle, K. Demuynck, J.F. Gemmeke, J.R. Bellegarda, and S. Sundaram, "Exemplar-Based Processing for Speech Recognition: An Overview," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 98–113, 2012.

[4] H. Hermansky and S. Sharma, "TRAPs – Classifiers of Temporal Patterns," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Sydney, Australia, 1998, pp. 1003–1006.

[5] H. Van hamme, "HAC-models: a Novel Approach to Continuous Speech Recognition," in *Proceedings of INTERSPEECH*, Brisbane, Australia, 2008, pp. 2554–2557.

[6] O. Räsänen, "A computational model of word segmentation from continuous speech using transitional probabilities of atomic acoustic events," *Cognition*, vol. 120, no. 2, pp. 149–176, 2011.

[7] K. Kinoshita, M. Souden, M. Delcroix, and T. Nakatani, "Single Channel Dereverberation Using Example-Based Speech Enhancement with Uncertainty Decoding Technique," in *Proceedings of INTERSPEECH*, Florence, Italy, 2011, pp. 197–200.

[8] M. Delcroix, K. Kinoshita, T. Nakatani, S. Araki, A. Ogawa, T. Hori, S. Watanabe, M. Fujimoto, T. Yoshioka, T. Oba, Y. Kubo, M. Souden, S. Hahm, and A. Nakamura, "Speech Recognition in the Presence of Highly Non-stationary Noise Based on Spatial, Spectral and Temporal Speech/Noise Modeling Combined with Dynamic Variance Adaptation," in *Proceedings of 1st CHiME workshop*, Florence, Italy, 2011, pp. 12–17.

[9] A. Mohamed, G. Dahl, and G. Hinton, "Acoustic Modeling using Deep Belief Networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 14–22, 2012.

[10] J.F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based Sparse Representations for Noise Robust Automatic Speech Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2067–2080, 2011.

[11] F. Weninger, J. Geiger, M. Wöllmer, B. Schuller, and G. Rigoll, "The Munich 2011 CHiME Challenge Contribution: NMF-BLSTM Speech Enhancement and Recognition for Reverberated Multisource Environments," in *Proceedings of 1st CHiME workshop*, Florence, Italy, 2011, pp. 24–29.

[12] A. Hurmalainen, J.F. Gemmeke, and T. Virtanen, "Modelling non-stationary noise with spectral factorisation in automatic speech recognition," *Computer Speech and Language*, vol. 27, no. 3, pp. 763–779, 2013.

[13] E. Yılmaz, J.F. Gemmeke, D. Van Compernolle, and H. Van hamme, "Noise-robust Digit Recognition with Exemplar-based Sparse Representations of Variable Length," in *IEEE Workshop on Machine Learning for Signal Processing (MLSP)*, Santander, Spain, 2012.

[14] D. Wang and J. Tejedor, "Heterogeneous Convolutive Non-Negative Sparse Coding," in *Proceedings of INTERSPEECH*, Portland, Oregon, USA, 2012.

[15] J. Barker, E. Vincent, N. Ma, C. Christensen, and P. Green, "The PASCAL CHiME Speech Separation and Recognition Challenge," *Computer Speech and Language*, vol. 27, no. 3, pp. 621–633, 2013.

[16] P. Smaragdis, "Convolutive Speech Bases and their Application to Supervised Speech Separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 1–14, 2007.

[17] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[18] A. Hurmalainen, J.F. Gemmeke, and T. Virtanen, "Compact Long Context Spectral Factorisation Models for Noise Robust Recognition of Medium Vocabulary Speech," in *Proceedings of 2nd CHiME workshop*, Vancouver, Canada, 2013, pp. 13–18.

Publication P8

A. Hurmalainen, J. F. Gemmeke, and T. Virtanen, “Compact Long Context Spectral Factorisation Models for Noise Robust Recognition of Medium Vocabulary Speech”, in *Proceedings of the 2nd International Workshop on Machine Listening in Multisource Environments (CHiME)*, Vancouver, Canada, 1. June 2013, pp. 13–18.

COMPACT LONG CONTEXT SPECTRAL FACTORISATION MODELS FOR NOISE ROBUST RECOGNITION OF MEDIUM VOCABULARY SPEECH

Antti Hurmalainen* *Jort F. Gemmeke*† *Tuomas Virtanen**

* Department of Signal Processing, Tampere University of Technology, Tampere, Finland
† Department ESAT, Katholieke Universiteit Leuven, Belgium

ABSTRACT

In environments containing multiple non-stationary sound sources, it becomes increasingly difficult to recognise speech from its short-time spectra alone. Long-context speech and noise models, where phonetic patterns and noise events may span hundreds of milliseconds, have been found beneficial in such separation tasks. Thus far the majority of work employing non-negative matrix factorisation to long-context spectrogram separation has been conducted on small vocabulary tasks by exploiting large speech and noise dictionaries containing thousands of atoms. In this work we study whether the previously proposed factorisation methods are applicable to more natural speech and limited noise context while keeping the model sizes practically feasible. Results are evaluated on the WSJ0 5k -based 2nd CHiME Challenge Track 2 corpus, where we achieve approximately 4% absolute improvement in speech recognition rates compared to baseline using the proposed enhancement framework.

***Index Terms*—** Spectral factorisation, speech recognition, noise robustness

1. INTRODUCTION

In conventional automatic speech recognition (ASR) it is common to employ short-term spectral features as the input for back-end recognition. A typical choice is computing mel-frequency cepstral coefficients (MFCCs) from 25 ms frames with a 10 ms shift. Hidden Markov models (HMMs), used to model temporal progression of speech, search for most likely paths by observing transition probabilities between two consecutive frames. Such short-term evaluation has been found sufficient for clearly spoken speech in optimal conditions. However, real-world speech recognition tasks rarely meet these expectations.

Apart from the linguistic variation taking place in casual speech, a major challenge for practical ASR is coping with signals corrupted by recording hardware, transmission channels, and environmental noise. The latter can be divided further into competing sources and acoustic phenomena such as reverberation. Whereas many kinds of constant channel errors and the effect of acoustic environment can be addressed with static compensation methods, additive noise from varying sources forms a greater obstacle. There is almost infinite variation in the sounds encountered in everyday situations, including semi-stationary background noise, sudden impacts, longer noise events, and competing speech. Especially the last example illustrates how the spectro-temporal behaviour of noise sources can be very close to actual target speech. Furthermore, in conditions falling below a 0 dB signal-to-noise ratio (SNR), noise sources start to dominate several spectral regions, making the short-time spectrum unreliable as a feature space for classification.

It has been demonstrated that increasing the temporal context of modelling units and observation windows is beneficial for discovering spectro-temporal regions dominated by speech or noise. Context of a few hundred milliseconds has been found relevant for speech modelling and perception in statistical speech analysis [1], intelligibility measurement [2] and direct observation of the auditory cortex [3]. The significance of temporal context for robust ASR has received further support in additive multi-source modelling with spectrogram factorisation, where the best results have been achieved by using observation windows spanning 200–300 ms [4, 5, 6].

However, an inherent downside of context expansion is that the modelling units become more specialised, and more units are required to cover the same event space than using a shorter context. In the previously referred experiments and related work, separation and classification quality were found to improve by using thousands of atoms even for small vocabulary tasks like 11-word Aurora-2 [7] and 51-word GRID/CHiME [6]. While early experiments have been conducted on large vocabulary, it is not clear whether the approach is viable for such tasks and eventually real world use.

To address this concern, we propose incorporating refined modelling methods to our non-negative matrix factorisation (NMF) framework. We apply long-context NMF to WSJ0-based 2nd CHiME Challenge Track 2 data, where medium vocabulary speech must be recognised from noisy mixtures ranging from +9 to -6 dB SNR. The identity of the target speaker is also unknown, which was not the case in the 1st CHiME Challenge involving difficult noise conditions [8]. New methods aiming at considerable basis reduction are compared to baseline results and large basis factorisation. In Section 2 we give the basics of spectrogram factorisation. Section 3 introduces recent methods which help in model size reduction. The experimental set-up is described in Section 4, whereafter results are listed and discussed in Section 5. Finally we present conclusions and ideas for future work in Section 6.

2. SPECTROGRAM FACTORISATION

By spectrogram factorisation we refer to techniques, where sound sources are separated in spectral domain by factoring a spectrogram matrix into its constituent parts. Furthermore, we concentrate on algorithms which take into account the temporal continuity of signals, that is, observe a context larger than individual frames. In earlier work, promising results have been achieved by using *non-negative* modelling. The motivation is that DFT resolution spectral magnitudes and features derived from them are mostly additive, thus non-negative additive models produce a good estimate of source component contribution.

Tuomas Virtanen has been funded by the Academy of Finland, grant #258708. The research of Jort F. Gemmeke was funded by IWT-SBO project ALADIN contract 100049.

A common characteristic in previously proposed work is that spectral modelling units and observation windows consist of T consecutive frames. A single spectrogram model, *atom*, is a $B \times T$ matrix, where B is the number of spectral bands in the feature space. Within a similarly sized observation window, the observed spectrogram $\mathbf{Y}$ is modelled as a sum

$$\mathbf{\Psi} = \sum_{l=1}^{L} x_l \mathbf{A}_l, \tag{1}$$

where $\mathbf{\Psi}$ is the estimate of $\mathbf{Y}$, L is the number of atoms (indexed by l), $\mathbf{A}$s are atom spectrograms, and xs are their *activation weights*. All spectral features and activation weights are non-negative. By assigning atoms into individual sources, in this case speech and noise, it is possible to derive single source estimates such as $\mathbf{\Psi}^{\mathrm{s}}$ for speech and $\mathbf{\Psi}^{\mathrm{n}}$ for noise by only including the chosen set's atoms in summing. These estimates are then employed to separate the original spectrogram into its components.

As the duration of an utterance, here denoted by T_{utt} frames, is generally longer than an atom, we need a model to represent the whole $B \times T_{\text{utt}}$ spectrogram as atom activations over time. Two alternative models have been used extensively in earlier work:

1. A 'sliding window' method, where $W = T_{\text{utt}} - T + 1$ overlapping $B \times T$ windows are extracted from $\mathbf{Y}$ in 1 frame steps, and factored individually [4]. The utterance spectrogram estimate $\mathbf{\Psi}$ is produced by averaging over window estimates, hence as an average of up to T single-window factorisations per frame. As atom and observation spectrograms can be vectorised and $\mathbf{X}$ solved from equation $\mathbf{\Psi} = \mathbf{AX}$, where $\mathbf{\Psi}$ is $BT \times W$, $\mathbf{A}$ is $BT \times L$ and $\mathbf{X}$ is $L \times W$, we call the method simply non-negative matrix factorisation (NMF) for short.
2. Non-negative matrix deconvolution (NMD), alternatively called convolutive NMF (CNMF), where the crucial difference to previously described NMF is that the utterance spectrogram estimate $\mathbf{\Psi}$ is produced jointly by all $\mathbf{X}$ entries via convolutive reconstruction. No averaging takes place as the overall spectrogram is a direct sum of timed activations.

Iterative update rules for determining $\mathbf{X}$ and $\mathbf{A}$ matrices are presented in detail in literature [9] and earlier work [4, 6]. Previous experiments suggest that sliding window NMF has inherent robustness against occasional mismatches and incorrect classification due to its averaging, whereas NMD is better suited for small atom count factorisation as its temporal model requires fewer shifted variants of each sound event than NMF. Both models are considered in this work with the focus being on NMD model reduction.

3. MODEL SIZE REDUCTION METHODS FOR FACTORISATION OF NOISY SPEECH

The basis generation algorithms in previously cited works have often relied on pseudo-random sampling of large amounts of *exemplars* from training material or from the noise neighbourhood of utterances to be recognised. The assumption is that given enough examples of sources, most observed events can be modelled as their linear combination. For abundant training data and model size, random sampling was found as good as initial attempts of refined selection. Later we have proposed informed speech basis reduction, replacing exemplars with state-centric templates, and noise basis reduction by NMD modelling [6]. Still, constraints such as small vocabulary, simplified grammar, or plentiful noise context were typically exploited in the experiments. In this section we present alternative speech and noise modelling methods, which produce compact bases for medium vocabulary speech separation in difficult conditions.

3.1. Variable length atoms

The first recently introduced model extension allows the length of atoms to vary within a basis. While in sliding window NMF the atom duration T is practically forced by design to be a constant in any single factorisation task, the same restriction does not apply to NMD. By using variable atom length it is possible to exploit long context and its benefits in separation whenever suitable, while also maintaining shorter units which also appear in natural speech and noise. Early experiments have been conducted on variable length bases for two-speaker separation [10] and robust ASR for small vocabulary [11], but the work presented here is among the first examples of variable length NMD modelling in semi-realistic ASR.

The convolutive utterance re-estimation formula for variable atom length T_l becomes

$$\mathbf{\Psi} = \sum_{l=1}^{L} \sum_{t=1}^{T_l} \mathbf{A}_{l,t} \overset{\rightarrow(t-1)}{\mathbf{X}_l}. \tag{2}$$

$\mathbf{A}_{l,t}$ is the t^{th} frame column vector of atom l, $\mathbf{X}_l$ is the l^{th} row vector of $\mathbf{X}$, and operator → shifts it right by $t-1$ columns.

In this work we use strongly variable-length speech bases by employing a basis acquisition algorithm similar to the one presented for CHiME/GRID speech data [11]. The algorithm starts from the longest permitted atom length $T = T_{max}$, inspects the speech training data, and attempts to find length T segments matching to each other. The measure used for match-finding is a combination of spectral data and monophone annotations to take into account both spectral and linguistic similarity. If a sufficiently large group of matching segments (here called a *cluster*) is found, a speech atom is formed by averaging the matching spectrograms. The corresponding areas of training data are flagged as taken. Thereafter the algorithm continues searching for clusters, reducing the segment length by one whenever the minimum cluster size requirement can no longer be met at current length T. Consequently a basis of template atoms is generated in a decreasing order of atom length and frequency of occurrence in the training data.

3.2. Multi-stage factorisation with speaker-dependent bases

In WSJ0-based CHiME Track 2 data, training and test speaker identities form disjoint sets. In other words, no exactly matching speaker model can be chosen for test factorisation, and no clues about test speaker characteristics are initially available. However, it is obvious that factorisation with a closely matching speaker model has a better chance of capturing correct speech features among noises which may include competing non-target speakers. Earlier it has been illustrated how NMD can act as a speaker identifier, when multiple speaker-dependent bases are used for factorisation and the relative activation weights of each speaker's atoms are observed [12].

Based on these findings, we propose a method which allows approximate speaker identification and basis selection by using multi-stage factorisation. In the initial stage, a small number of atoms from all training speakers are used, and relatively few NMD iterations are computed. In each subsequent stage, speaker activity weights are used for selecting the best matching bases, while more atoms from the chosen speakers are introduced to factorisation. Eventually the system will converge to a small set of training speakers, whose speech profiles match best to the target speaker. The details for the presented set-up are given in Section 4.4. By dynamic management

Figure 1: Histogram of speech atom lengths from variable length basis acquisition, ranging from 8 to 50 frames.

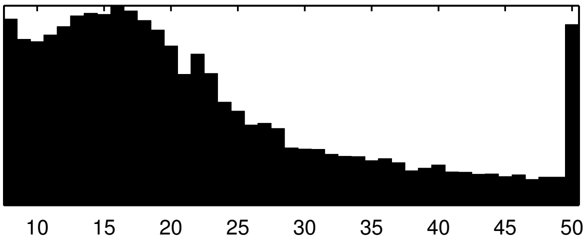


Table 1: Statistics of speech bases used during multi-stage factorisation of the CHiME Track 2 evaluation set. For each stage, the number of active speaker bases and their combined atom count is reported as minimum, mean and maximum values.

Stage	Speakers			Atoms		
	min	mean	max	min	mean	max
1	83	83	83	4150	4150	4150
2	9	24.3	36	900	2427	3600
3	2	8.9	17	612	3023	5754
4	1	3.8	9	304	1305	3246

of the number and size of bases, it is possible to perform multi-speaker modelling and semi-matched final factorisation, while the amount of simultaneously active atoms remains low. For accelerated basis set reduction, we use a group sparsity constraint, which favours solutions where activations come from a small number of bases [12].

3.3. Pre- and online-adaptation of noise atoms

For acquisition of noise models, there are three important sources whose availability and significance depends on the recognition task. First, we may have fixed training material for pure noise. Second, there is a varying amount of noise context surrounding target speech. Finally, noise can be estimated from the utterance itself by capturing features which do not match to any speech models. In previous work, all three methods have been exploited [6, 7, 13, 14] with occasional further extensions such as artificial noise atoms [15].

Previously we have achieved the best results by sampling large exemplar bases randomly from training data [4] or semi-randomly from the local context [6] according to availability. However, both methods are prone to including a lot of redundancy or unnecessary, near-silent spectral data. Furthermore, exemplars sampled from additive multi-source mixtures cannot model accurately the same events appearing alone or in different combinations. Therefore in this work we use methods based on NMD learning to acquire smaller noise models with a higher efficiency.

Regardless of which data is used for noise learning, we apply iterative NMD atom update rules described in literature [9, 16]. For CHiME Track 2 data, we use two sources for noise atoms: first, background training data which is first reduced to its loudest sections, and second, the ‘embedded’ utterances with 5 seconds of noise context before and after. It has been found that to prevent overfitting and fragmentation of learnt atoms into unusably small spectro-temporal units, adaptation should be terminated earlier than the commonly employed amount of factorisation iterations for fixed bases. Computationally the simplest way to implement this is to reduce the number of iterations to approximately 20–30 (compared to 200–400 of earlier work), which can be achieved in long semi-supervised factorisation by only performing a basis update after a certain interval of activation update iterations.

4. EXPERIMENTAL SET-UP

A factorisation framework was designed for the 2nd CHiME Challenge medium vocabulary (Track 2) dataset [17]. Its speech data consists of WSJ0 5k vocabulary utterances and is divided as follows:

- 7138 training utterances jointly from 83 speakers, both ‘clean’ (without additive noise) and mixed at a random SNR
- 409 development test utterances jointly from other 10 speakers and repeated at 6 SNRs
- 330 evaluation test utterances from other 8 speakers, 6 SNRs

Noisy utterances are mixed with non-stationary multi-source household noise at SNRs ranging from +9 to -6 dB in 3 dB steps. Noise data contains natural room reverberation. For speech data, similar impulse responses are simulated. All utterances are available with 5 seconds of noise context before and after the utterance. Approximately seven hours of pure noise data is also available for training. Recognition is measured by HTK toolkit’s ‘Err’ word error rate.

4.1. Feature space

All factorisation experiments were conducted in monaural 40-band mel-spectral magnitude space. Features were extracted from binaural input signals with a frame length of 25 ms and frame shift of 10 ms, and averaged in absolute magnitude value domain. Mel bands were reweighted by a fixed equalisation curve derived from 2-normalisation of noisy 0 dB training utterances.

4.2. Speech bases

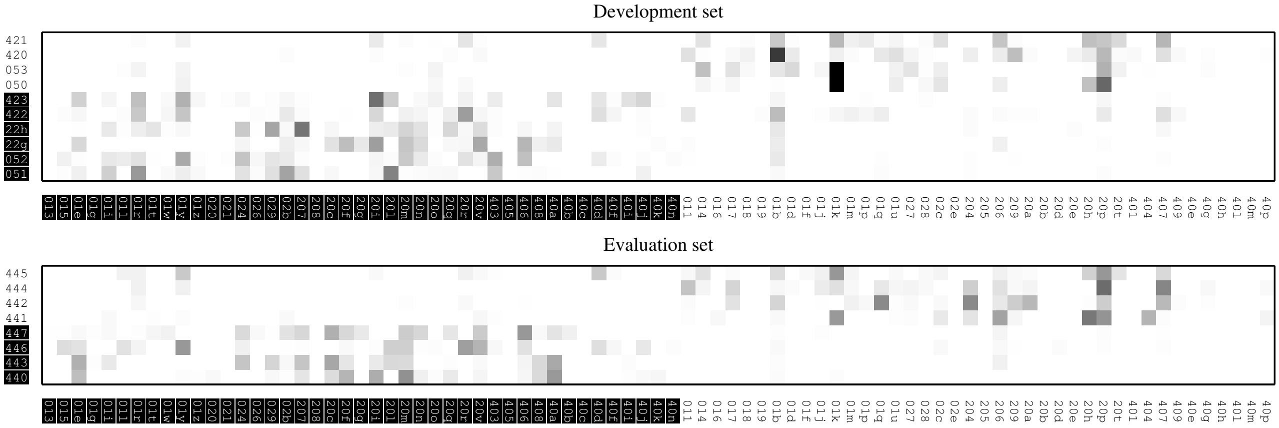
A variable-length speech basis was generated for each training speaker similarly to the algorithm described for 1st CHiME challenge data [11]. The similarity measure for frame vectors consisted of dot product between normalised, square root compressed mel magnitudes augmented with delta features, and monophone labels acquired from forced alignment using the baseline recogniser. Similarity between frames i and j was computed as

$$c_m(i,j) = c_\mathrm{s}(i,j) + c_l(i,j), \tag{3}$$

where the merged similarity c_m is the sum of spectral vector dot product c_s and correlation of monophone labels c_l, the latter ranging from 0 to 0.06 depending on how closely monophones and their substates matched in annotations. Sequences where all mutual frame pairs produced at least 0.92 total similarity were considered for clustering. A cluster was selected for atom construction if its source segments covered at least 0.15% of the speaker’s noiseless training material. In other words, long segments were allowed to form atoms with fewer matches than short segments. Atom lengths ranged from 46 to 4 in clustering, whereafter the 2 preceding and following frames were added to atoms as their content is implied by delta features. Consequently the final length of speech atoms was between 8 and 50 frames (80–500 ms).

Figure 1 illustrates the distribution of speech atom lengths in combined speech bases. We notice that large variation takes place, reflecting the multitude of phonetic unit lengths appearing in natural speech. A large peak can be seen at length 50. Even longer correlating segments could be found, but their value for factorisation

Figure 2: Similarity of test speakers (y-axis) to training speakers (x-axis), measured as the amount of speaker-dependent basis activations in the last stage of 9 dB test set factorisation. For each test identity, the sum of activations is normalised to unity. Similarity increases toward black with the maximum intensity being 0.4. White-on-black identity names belong to male speakers, black-on-white to females.



becomes negligible thus they were truncated to the chosen maximum value. Mean atom length was 22.2 frames, approximately matching the previously favoured fixed contexts of 200–300 ms [4, 6, 13].

The number of atoms in the 83 speaker-dependent bases was from 276 to 397 with a mean value of 344 and combined atom count of 28579. Speakers with more variable pronunciation generated larger bases than very consistent ones. Because the cluster size was defined as a percentage of available data, no notable difference was present between speakers with fewer or more training utterances. By comparing the basis sizes to the 5000 word vocabulary, it is clear that the typical unit modelled was shorter than a complete word.

4.3. Noise bases

Two noise modelling methods were used: a fixed noise basis acquired by NMD learning from background training material, and online-adapted noise model from the embedded utterances.

For fixed basis acquisition, the seven-hour training material was first reduced to its loudest 20% frames, measured by spectral magnitudes. From the remaining material, segments shorter than 5 frames were removed, while the rest were padded by 10 frames before and 30 after, approximating the usual temporal decay profile of noise events. Thereafter the segments were faded in and out with a 10-frame transition, and concatenated into approximately five minute blocks of significant noise events. Each block was factored with 25 iterations of NMD basis adaptation to produce atoms with a joint duration of 10% of block length, that is, approximately 60 atoms of length 50 frames per block. While no attempt was made to force noise atoms into shorter or variable duration, in practice this often happened due to some of the atoms modelling short-duration noise events. The procedure as a whole generated 1729 fixed noise atoms.

Direct adaptation of noise atoms from embedded utterances followed mostly similar principles, yet employed significantly fewer atoms. The details are described in the next subsection.

4.4. Multi-stage factorisation

Training and test file factorisation was conducted using the 'embedded' files with 5 seconds of noise context to both directions. After feature extraction, the following bases were set up:

- Speech bases: for test files all 83 speaker-dependent bases, for training files all except self
- A variable amount of randomly initialised adaptive noise atoms, enough to cover 75% of embedded utterance duration
- Optionally, the fixed 1729-atom noise basis (See section 4.3.)

The motivation for given speech basis choices was to use a set of bases disjoint from the target identity. For development and evaluation sets this was automatically the case. For training utterances, the true matching identity was left out to prevent oracle modelling.

The adaptive noise atom count was left slightly below the amount required to cover all embedded utterance frames in order to promote discovery of recurrent features. These atoms were readapted from scratch for each utterance from its own context alone. Training and evaluation were run with and without the fixed noise basis to study whether the methods are applicable to entirely new situations where pre-training of noise models is not an option.

For factorisation, variable-length NMD was used with generalised Kullback-Leibler divergence as the spectral distance measure, and L_1 penalty as the sparsity constraint similarly to earlier work. L_1/L_2 group sparsity penalty was induced on speech activations as presented previously [12], with each speaker's atoms forming a group. Sparsity weights were defined by brief experimentation on development utterances and set to 0.07, 0.1, 0.1 and 0.11 for speech, groups, adaptive noise and fixed noise (respectively) when the latter was used, and 0.08, 0.1 and 0.1 for the rest when not. All sparsity values are proportional to the mean value of basis atom 1-norms.

Factorisation had four stages with basis pruning as follows:

1. All speech bases, 50 atoms per speaker, 50 iterations
2. Reduced set of bases, 100 atoms per speaker, 50 iterations
3. Further reduced set of bases, all atoms, 100 iterations
4. Final reduced set of bases, all atoms, 100 iterations

Each partial basis consisted of the first (longest) atoms of complete speaker-dependent bases. Between stages, activation matrix sums were calculated for each speaker dependent basis. A threshold value was set 10–20% from the geometric mean toward the largest value to remove all except the best matching identities. Activation weights of remaining speech atoms were left as is, whereas newly introduced atoms were given a small initial weight of 0.001. Noise atoms or their activation were not changed between stages.

Table 2: Results for CHiME Track 2 noise robust speech recognition, listed as word error rate ('Err') over SNRs. Tables on the left and the right show results for development and evaluation sets, respectively. First, the baseline results using provided 'noise' models are given. The next lines show results for proposed enhancement using adaptive noise atoms only, and then for both adaptive and fixed noise atoms. Finally reference results for large basis NMF are shown. Results are evaluated using provided and re-trained GMMs.

SNR (dB)		9	6	3	0	-3	-6	avg
Baseline ('noise')		44.34	49.05	55.71	59.89	67.43	73.17	58.27
Adapt. only	noise	44.03	48.91	55.04	58.35	65.97	71.19	57.25
	re-trained	42.93	48.24	53.76	57.53	64.71	70.92	56.35
Adapt. +fixed	noise	44.19	47.25	53.27	56.53	63.93	69.47	55.77
	re-trained	42.28	45.54	51.45	55.43	62.61	69.26	54.43
Large NMF	noise	43.33	46.75	51.66	56.51	64.61	69.28	55.36
	re-trained	39.13	44.18	47.65	52.29	60.56	66.23	51.67

(a) Development set

SNR (dB)		9	6	3	0	-3	-6	avg
Baseline ('noise')		41.73	45.32	51.06	58.42	63.09	70.43	55.01
Adapt. only	noise	42.59	45.19	49.71	56.53	61.76	66.75	53.76
	re-trained	40.30	44.44	48.70	54.04	60.34	66.90	52.45
Adapt. +fixed	noise	41.60	44.16	50.29	54.80	60.34	66.34	52.92
	re-trained	38.76	41.53	47.99	51.73	58.83	66.71	50.93
Large NMF	noise	42.35	44.35	48.81	54.01	60.17	65.18	52.48
	re-trained	37.40	39.14	43.51	50.94	55.58	61.85	48.07

(b) Evaluation set

Basic statistics of basis and atom counts in each stage are listed in Table 1 for the test set (with the fixed noise basis enabled). Notably, the simultaneous speech atom count never exceeded 5754, and the last stage employed on average 3.8 bases and 1305 atoms.

Figure 2 illustrates the convergence of different test speakers' (y-axis) factorisation toward matching training speaker bases (x-axis). 9 dB SNR experiments were used for the plot to minimise noise interference. We can observe that even though approximately 40 different utterances were factorised per test speaker, the algorithm generally converged toward a spiky distribution of only a few matching bases. The bases were also mostly from the same gender as the test speaker, and the set was unique for each individual speaker. Comparison by listening confirmed that approximately similar speaker profiles were generally found.

Speech and fixed noise activations were only permitted in the actual utterance area, whereas adaptive noise activations were permitted also in the noisy context to capture the immediate noise environment. As the adaptive basis size was generally below 30 atoms and only updated every 10 iterations (of total 300), factorisation effort was mostly concentrated on the noisy speech, and the overall complexity of the system remained comparable to previous small vocabulary experiments.

4.5. Enhancement and recognition

The activation matrices acquired from NMD were used to generate speech and noise spectrogram estimates as described in Sections 2 and 3.1. Mel spectrograms were mapped back to linear frequency domain and used as a time-varying filter defined as $\mathbf{\Psi}^{\mathrm{s}}/(\mathbf{\Psi}^{\mathrm{s}} + \mathbf{\Psi}^{\mathrm{n}})$ for the original noisy spectrograms [6].

Because the sparse NMD model with adaptive atoms occasionally produces rapidly changing spectro-temporal behaviour with heavy filtering in fully masked segments, it was found beneficial to apply a 0.1 minimum value to the filter weight value normally ranging from 0 to 1. Enhanced signals were recognised using the CHiME HTK tools, both with the multi-condition noise trained baseline models and models re-trained with enhanced training data.

For comparison, we also implemented a sliding window NMF system employing considerably larger exemplar bases similarly to earlier work. 10000 speech exemplars and 4000 noise exemplars were sampled randomly from training material, whereafter approximately 1000 noise exemplars were added from the context. Feature space, factorisation and enhancement followed generally similar principles to those presented for Aurora-2 and 1st CHiME data [4, 6], and for applicable parts they matched the NMD setup.

5. RESULTS AND DISCUSSION

Results for speech recognition experiments are given in Table 2 as word error rates (HTK 'Err') per SNR, separately for development and evaluation sets. The first row shows results using baseline 'noise' models and unenhanced waves. The next rows list results for proposed enhancement using adaptive noise only, and for adaptive+fixed noise. The last rows list results for reference NMF enhancement using large exemplar bases. Enhanced signals were evaluated using the baseline 'noise' models, and with GMMs re-trained from matching training set enhancement.

We observe that enhancement with the proposed approach generally yields improvement over the baseline already on the standard back-end models. Expectedly including a fixed noise basis acquired from background training material provides further improvement over just using noise adaptation from the embedded utterance. Without back-end re-training, the proposed system with both noise models is approximately comparable (2–3 % over baseline) to NMF with large exemplar bases. In re-training, the gap increases so that the improvements over unenhanced baseline are approximately 4% and 7% for proposed and NMF factorisation, respectively.

The proposed framework is our first attempt to develop a relatively lightweight factorisation and enhancement system for medium-vocabulary speech recognition in difficult conditions. Compared to the GRID-based 1st CHiME set [8], the new WSJ-based corpus introduced several new challenges. The 5000 word vocabulary with only limited training data available for each speaker requires a different approach to generating speaker-dependent speech bases. Furthermore, test identities coming from disjoint speaker sets prevented selecting a perfectly matching speech model.

We investigated using several small speaker-dependent bases, which complement each other concerning both vocabulary and speaker characteristics. A clear benefit of (approximate) identity matching is the ability to separate a target speaker from competing speakers, which is more difficult with a speaker-independent basis modelling all speakers simultaneously. From Figure 2 we see that at least at high SNRs the algorithm was able to find similar speaker profiles. An obvious problem of the method is that non-target speakers have a good chance of activating an alternative set of bases, and at < 0 dB even dominating the selection process. Currently this is only prevented by vocabulary matching via long context atoms. Further methods for correct selection could include spatial estimation and preliminary decoding during the selection process.

In noise modelling, initial results suggest that a noise model adapted from a 5 second context only has a limited separation capability. Acquiring a comprehensive model beforehand improves results significantly. However, the obvious problem is applying the method to new noise environments. In a real-world system, continuous noise model updating during pauses in speech would be preferable in order to maintain a maximally good match. Such a system for continuous NMD recognition has already been proposed [18].

With respect to model complexity and the goal of achieving feasible basis sizes, we can observe that the proposed framework managed to improve average speech recognition rates by approximately 4% (absolute) compared to the unenhanced baseline with an average basis size of 1305 final stage speech atoms, 1729 fixed noise atoms, and generally less than 30 adaptive noise atoms – approximately $1/5^{th}$ of the reference NMF basis size. While more atoms were temporarily used for speaker selection, it must be noted that in these experiments we always started from all 83 candidates for each utterance. In practice, there is a lot of redundancy among the models with some of them barely activating at all, and in real world it rarely applies that speaker adaptation should be repeatedly started from scratch. Therefore we expect that the multi-speaker basis sizes could be easily reduced further. Regarding vocabulary size, already the current bases modelled sub-word units of a vocabulary 15 times larger than average atom count and covered a large part of common linguistic units, hence the requirements for truly large vocabulary should not be considerably greater.

6. CONCLUSIONS

We presented a spectrogram factorisation framework designed for medium vocabulary speech recognition using long temporal context yet compact bases. Several emerging or wholly novel ideas were proposed, including variable length modelling, multi-stage factorisation with basis pruning, and two noise models used in conjugation.

With refined bases, it was found feasible to separate unknown speaker's speech from very noisy mixtures with models smaller than were previously used for small vocabulary tasks with matching speaker identity. Approximately 4% absolute reduction was obtained in average word error rate in evaluation on the 2nd CHiME Challenge Track 2 corpus. As several novel aspects were introduced and combined for a new task with limited parameter tuning, we expect further improvements when their standalone and interoperation characteristics becomes better understood. Nevertheless, already the initial results appear promising regarding robust real-world speech recognition with practically applicable factorisation model sizes.

7. REFERENCES

[1] O. Räsänen and U. K. Laine, "A method for noise-robust context-aware pattern discovery and recognition from categorical sequences," *Pattern Recognition*, vol. 45, no. 1, pp. 606–616, 2012.

[2] T. M. Elliott and F. E. Frédéric, "The Modulation Transfer Function for Speech Intelligibility," *PLoS Computational Biololgy*, vol. 5, no. 3, pp. e1000302, 2009.

[3] B. N. Pasley, S. V. David, N. Mesgarani, A. Flinker, S. A. Shamma, N. E. Crone, R. T. Knight, and E. F. Chang, "Reconstructing Speech from Human Auditory Cortex," *PLoS Biology*, vol. 10, no. 1, pp. e1001251, 2012.

[4] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based Sparse Representations for Noise Robust Automatic Speech Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2067–2080, 2011.

[5] F. Weninger, M. Wöllmer, J. Geiger, B. Schuller, J. F. Gemmeke, A. Hurmalainen, T. Virtanen, and G. Rigoll, "Non-negative Matrix Factorization for Highly Noise-robust ASR: To Enhance or to Recognize?," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Kyoto, Japan, 2012, pp. 4681–4684.

[6] A. Hurmalainen, J. F. Gemmeke, and T. Virtanen, "Modelling Non-stationary Noise with Spectral Factorisation in Automatic Speech Recognition," *Computer Speech and Language*, vol. 27, no. 3, pp. 763–779, 2013.

[7] J. F. Gemmeke, A. Hurmalainen, T. Virtanen, and Y. Sun, "Toward a Practical Implementation of Exemplar-Based Noise Robust ASR," in *Proceedings of European Signal Processing Conference (EUSIPCO)*, Barcelona, Spain, 2011, pp. 1490–1494.

[8] J. Barker, E. Vincent, N. Ma, C. Christensen, and P. Green, "The PASCAL CHiME Speech Separation and Recognition Challenge," *Computer Speech and Language*, vol. 27, no. 3, pp. 621–633, 2013.

[9] A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari, *Nonnegative Matrix and Tensor Factorizations*, Wiley, 2009.

[10] D. Wang and J. Tejedor, "Heterogeneous Convolutive Non-Negative Sparse Coding," in *Proceedings of INTERSPEECH*, Portland, Oregon, USA, 2012.

[11] A. Hurmalainen and T. Virtanen, "Acquiring Variable Length Speech Bases for Factorisation-Based Noise Robust Automatic Speech Recognition," (to be published).

[12] A. Hurmalainen, R. Saeidi, and T. Virtanen, "Group Sparsity for Speaker Identity Discrimination in Factorisation-based Speech Recognition," in *Proceedings of INTERSPEECH*, Portland, Oregon, USA, 2012.

[13] F. Weninger, J. Geiger, M. Wöllmer, B. Schuller, and G. Rigoll, "The Munich 2011 CHiME Challenge Contribution: NMF-BLSTM Speech Enhancement and Recognition for Reverberated Multisource Environments," in *Proceedings of International Workshop on Machine Listening in Multisource Environments (CHiME)*, Florence, Italy, 2011, pp. 24–29.

[14] R. Vipperla, S. Bozonnet, D. Wang, and N. Evans, "Robust Speech Recognition in Multi-Source Noise Environments using Convolutive Non-Negative Matrix Factorization," in *Proceedings of CHiME workshop*, Florence, Italy, 2011, pp. 74–79.

[15] J. F. Gemmeke and T. Virtanen, "Artificial and Online Acquired Noise Dictionaries for Noise Robust ASR," in *Proceedings of INTERSPEECH*, Makuhari, Japan, 2012, pp. 2082–2085.

[16] P. Smaragdis, "Convolutive Speech Bases and their Application to Supervised Speech Separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 1–14, 2007.

[17] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The Second 'CHiME' Speech Separation and Recognition Challenge: Datasets, Tasks and Baselines," in *Proceedings of ICASSP*, Vancouver, Canada, 2013.

[18] A. Hurmalainen, J. F. Gemmeke, and T. Virtanen, "Detection, Separation and Recognition of Speech From Continuous Signals Using Spectral Factorisation," in *Proceedings of European Signal Processing Conference (EUSIPCO)*, Bucharest, Romania, 2012, pp. 2649–2653.

“To a mathematician, real life is a special case.”

Tampereen teknillinen yliopisto
PL 527
33101 Tampere

Tampere University of Technology
P.O.B. 527
FI-33101 Tampere, Finland

ISBN 978-952-15-3371-6
ISSN 1459-2045