LEARNING VOCAL MODE CLASSIFIERS FROM HETEROGENEOUS DATA SOURCES

Zhao Shuyang, Toni Heittola, Tuomas Virtanen

Tampere University of Technology
Signal Processing Department
korkeakoulunkatu 1, Tampere 33720, Finland
shuyang.zhao@tut.fi, toni.heittola@tut.fi, tuomas.virtanen@tut.fi

ABSTRACT

This paper targets on a generalized vocal mode classifier (speech/singing) that works on audio data from an arbitrary data source. Previous studies on sound classification are commonly based on cross-validation using a single dataset, without considering training-recognition mismatch. In our study, two experimental setups are used: matched training-recognition condition and mismatched training-recognition condition. In the matched condition setup, the classification performance is evaluated using cross-validation on TUT-vocal-2016. In the mismatched setup, the performance is evaluated using seven other datasets for training and TUT-vocal-2016 for testing. The experimental results demonstrate that the classification accuracy is much lower in mismatched condition (69.6%), compared to that in matched condition (95.5%). Various feature normalization methods were tested to improve the performance in the setup of mismatched training-recognition condition. The best performance (96.8%) was obtained using the proposed subdataset-wise normalization.

Index Terms: sound classification, vocal mode, heterogeneous data sources, feature normalization

1. INTRODUCTION

In this study, we aim at a generalized vocal mode (speech/singing) classifier, working on audio data from arbitrary sources. A generalized vocal mode classifier can potentially save a lot of time when finding interesting parts in a video, along with established vocal activity detection techniques [1, 2]. As an example, the singing part from a talent show episode can be easily found on YouTube.

The captured audio is affected by the recording device, acoustic space and background noises. The acoustic space and the recording device are collectively defined as transmission channel. In practice, the training-recognition mismatch is a significant problem: a classifier often fails when working on audio data captured using a different recording setup. However, majority of previous sound classification studies are based on a single dataset using cross-validation [3, 4, 5], without considering the cases of training-recognition mismatch. We call it a *homogeneous recognition scenario*, when training and testing data are from the same recording setup. We call it a *heterogeneous recognition scenario*, when recognition data is from different recording setups compared to the training data.

In previous studies, feature normalization has been shown effective to cope with training-recognition mismatch in robust speech recognition [6, 7, 8]. Mean-variance normalization (MVN) [9] scales features in each data source to have zero-mean and unit-variance. Histogram equalization (HE) [7, 8] aims at a more sophisticated matching over the histogram from a distribution basis to a distribution target. Notably, there is a significant difference between our study and robust speech recognition. Taking the experimental framework Aurora [10] used in [7, 8] as an example, a single clean speech dataset is used for training. The background noises of different environment are added to clean data to be used as testing material, thus the main mismatch is the background noises. In our study, the training material is from a few different datasets instead of one to cover various speech and singing styles. The main mismatch between the datasets is in channel effect instead of background noise, since all the datasets are recorded in relatively silent environment.

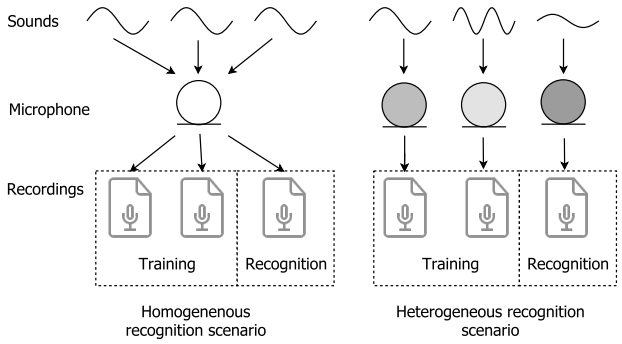

Figure 1: An example of a homogeneous recognition scenario and a heterogeneous recognition scenario.

This study deals with the training-recognition mismatch when learning vocal mode classifiers from heterogeneous data sources. Firstly, we investigate the difference in performance between homogeneous recognition scenario and heterogeneous recognition scenario. Secondly, we evaluate various feature normalization methods to improve the classification performance in heterogeneous recognition scenario. The main focus is the data scope to perform feature normalization, which is seldom investigated in previous studies. Besides the obvious recording-wise and dataset-wise normalization, subdataset-wise normalization is proposed. The normalization data scopes are evaluated along with MVN and HE. A new dataset TUT-vocal-2016 is introduced to evaluate the classification performance.

The organization of this paper is as follows. The method is described in Section 2. The used datasets are discussed in Section 3 and the experimental results for evaluation are given in Section 4. The conclusions are drawn in Section 5.

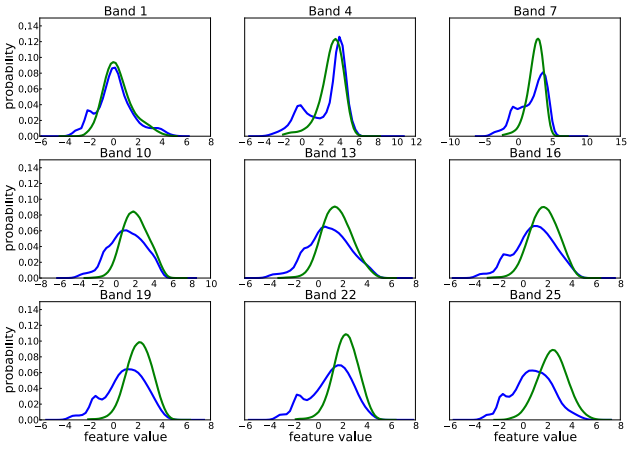

Figure 2: Feature distribution in CHiME2010 and Arctic dataset, illustrated in green and blue lines, respectively. The visualized features are log Mel-band energies from nine different bands.

2. METHOD

A vocal mode classifier takes an audio recording as input and the output is the predicted vocal mode corresponding to every second in the recording. The vocal mode classifier follows an established setup in the domain of sound classification: log-mel band energies as features and a multilayer perceptron as the classifier [11, 12]. In addition to the established setup, feature normalization is performed on the log-mel band energies.

2.1. Acoustic Features

The acoustic features are calculated as follows. The audio amplitude is normalized, scaling the maximum amplitude of a recording to one. The audio signal is divided into frames with length of 30 ms and 50% overlap. The number of Mel filter banks is 30, ranging from 25 Hz to 8000 Hz.

In order to investigate the difference in transmission channel between different data sources, the feature distributions of two speech datasets, CHiME2010 [13] and Arctic [14], are visualized in Figure 2. The histogram plots are obtained by dividing the interval [-4σ, 4σ] of each feature coefficient into 50 bins. Only features from non-silent frames are taken into account. As is shown in Figure 2, each feature coefficient in the CHiME2010 dataset is distributed around a single peak, similar to the normal distribution. In contrast, most coefficients in Arctic dataset are distributed around two peaks. Both CHiME2010 and Arctic are speech datasets containing balanced English utterances recorded in relatively silent environment, however the feature distributions are largely different, which reveals the the difference between the two datasets in terms of channel effect.

2.2. Feature normalization

A transmission channel introduces a time-invariant distortion to the original signal, under the assumption of linear system. It is assumed that there exists an invariant global distribution for voice signal, before transmitting through a channel [6]. If different recording setup is used, the global distribution becomes transformed. The global distribution using a recording setup can be estimated using available data from the source. Feature normalization aims at removing the noise and channel effect by matching the overall feature distributions of different data sources.

Two types of feature normalization techniques are considered: mean-variance normalization (MVN) [9] and quantile equalization (QE) [15]. They are simple and require not too much data from a data source to estimate the feature distribution, compared to more complicated and elaborated methods such as full histogram equalization [7], feature space rotation [16] and vocal tract length normalization [17].

In practice, it is quite often unknown what recordings are from the same recording setup. The audio data inside a recording is surely homogeneous, however the amount of data in a single piece of recording may not be sufficient to estimate the feature distribution of the source. Another solution is dataset-wise normalization, based on the assumption that the audio in the same dataset is recorded under very similar condition. However, this is not always a valid assumption. As an example, some audio datasets are collected in parallel using different recording devices in different environment. We use a term *data scope*, within which the feature distribution is estimated and feature normalization is performed. Global normalization, as a reference, scales all the data the same way, based on the statistics of the whole training material.

In addition, we propose another approach, where datasets are decomposed into sub-datasets based on K-means clustering on recordings. The number of clusters is defined proportionally to the data amount, with two hours of non-silent material in the dataset corresponding to one cluster.

Overall, we consider two feature normalization techniques and three normalization data scopes. In mean-variance normalization, a feature vector $\mathbf{x}$ in a data scope $\mathbf{X}$ is normalized as

$$\mathbf{x}_{norm} = \frac{\mathbf{x} - \mu}{\sigma}, \tag{1}$$

where μ and σ is the mean and standard deviation within the data scope $\mathbf{X}$.

Quantile equalization estimates a transformation function for each feature coefficient based on the quantile statistics of the data scope as basis and the whole training set as target. Five critical values: minimum, 25th-percentile, median, 75-percentile and the maximum are used to divide the range of a feature coefficient into four bins. The value of kth critical value for ith coefficient is denoted as Q_k^i and $\hat{Q}_k^i$, respectively for the basis and target distribution. A feature coefficient x^i is normalized as

$$x_{norm}^i = \hat{Q}_k^i + (x^i - Q_k^i)\frac{\hat{Q}_{k+1}^i - \hat{Q}_k^i}{Q_{k+1}^i - Q_k^i} \tag{2}$$
$$\forall x^i \in Q_k^i < x < Q_{k+1}^i.$$

2.3. Supervised learning

Multilayer perceptron (MLP) [18] is a basic type of artificial neural network, consisting of layers of nodes with each layer fully connected to the next one. Feature vectors are given to the network as input and the output corresponds to target classes. The implementation is based on Keras [19] using Theano [20] as backend.

Let us denote the node values of input layer as $\mathbf{h}^1 = \mathbf{x}$ and the node values of kth layer as $\mathbf{h}^k$. Given the node values of $k-1$th

Name	Class	Duration	Ref
CHiME	Speech	7h 06m	[13]
Arctic	Speech	6h 27m	[14]
CHAINS	Speech	2h 19m	[21]
Multitrack2013	Sing	17h 10m	[22]
Marl	Sing	1h 51m	[23]
Tonas Flamenco	Sing	0h 13m	[24]
TUT-VOX	Sing	0h 48m	-
TUT-vocal-2016	Both	3h 15m	-

Table 1: Datasets used in our experiments. Used length is the non-silent part of used recordings in a dataset. The duration is reported excluding silence.

layer, the node values of kth layer are calculated as

$$\mathbf{g}^{k-1} = \mathbf{W}^k \mathbf{h}^{k-1} + \mathbf{b}^k, 2 \leq k < M \quad (3)$$

$$\mathbf{h}^k = \mathcal{F}(\mathbf{g}^{k-1}). \quad (4)$$

Eq. (3) shows the linear transformation operation on $k-1$th layer of the neural network, where $\mathbf{W}^k \in \mathbb{R}^{S_{k-1} \times S_k}$ is a weight matrix between layer $k-1$ and layer k. S_k is number of neurons in layer k. A bias vector is denoted as $\mathbf{b}^k$. A non-linear activation function $\mathcal{F}$ is applied element-wise on the linear transformation outputs. The total number of hidden layers is two. Sigmoid function is used as activation function for hidden layers and the output layer ($\mathbf{h}^4 = \mathbf{y}$).

Context windowing is used for the neural network input: the consecutive feature frames $[\mathbf{x}[t-R], ..., \mathbf{x}[t], ..., \mathbf{x}[t+R]]$ are stacked together to form a single feature vector $\mathbf{x_c}[t]$ to represent temporal dynamics, where R is number of the past and future frames. We use $N_{cw} = 2R + 1 = 25$ to denote the total number of frames used in a context window. In order to smooth the neural network output, mean filter is used for neural network output as $[\mathbf{y}[t-L], ..., \mathbf{y}[t], ..., \mathbf{y}[t+L]]$. The size of the mean filter is $N_{mf} = 2L + 1 = 35$.

3. DATASETS

There is not any public dataset designed for speech/singing classification. However, there are many speech datasets designed for speech recognition and several singing datasets designed for music research. Three speech datasets and four singing datasets are selected as training material based on the variability and accessibility. The list of used datasets is shown in Table 1. In addition, we collected a new dataset TUT-vocal-2016 that contains both speech and singing to evaluate trained classifiers.

3.1. Datasets for training

All the speech datasets contain English speech from both male and female. CHiME dataset [13] contains speech utterances from 34 speakers with reverberation. Arctic dataset [14] is a clean speech dataset designed for speech synthesis and speech recognition, contributed by 7 speakers. CHAINS dataset is contributed by 36 speakers, including normal, fast and whispered speech. Only the normal speech and whispered speech utterances are used in this study.

Four singing databases are used. Multitrack2013 covers singing styles of pop and pop rock [22]. Tonas Flamenco [24] contains only Flamenco singing. The Marl dataset [23] contains pop singing and rap. Recordings containing rapping have been excluded in our experiments, since it is ambiguous if rapping belongs to speech or singing. TUT-VOX is a proprietary dataset containing acappella singing in English and Finnish.

3.2. TUT-vocal-2016

In order to make a proper evaluation for vocal mode classification, we introduce a new dataset TUT-vocal-2016. The core idea is to have audio where the same person is speaking and singing, preferably the same language content. The dataset is contributed by 20 volunteers, 10 females and 10 males. Each volunteer is required to choose four songs. The volunteer is required to sing from one to one half minutes of each song, thus all recordings weigh similarly in the evaluation. There are 80 pieces of singing collected, from a set of 21 different songs. The lyric of the songs is read out by each volunteer in three types: normal speech, whispered speech and shouted speech. The shouted speech is not used in this study since we have found very little shouted speech as training material.

3.3. Annotation

We use frame-level voice activity annotation, by which the silent parts in recordings are excluded for both training and testing. The frame-level activity annotation was obtained using two automatic approaches. The principle is to exclude all the silent frames in the evaluation and a small part of voices annotated as silence is tolerated.

Speech utterances were mostly short and contained usually only silence at the beginning and at the end of the signal. A simple energy-based scheme was chosen for this type of signals. In the scheme, 10% of the average RMS-energy was used as threshold to detected non-silence (active) segments. This scheme worked best with signals having mostly active segments and most of the energy is also concentrated in these vocal segments.

The acappella singing contains longer silent segments and in some cases added effects like reverberation making it hard to use such a simple threshold. For these type of signals, a binary classifier based approach was used [25]. In this approach, 10% of lowest energy frames within a recordings are used to train Off-class and 10% of highest energy frames is used to train On-class. The classifier was used to get probability of frame belonging to the On-class (active). Classification was done by defining the probability threshold as weighted mean between top 10% and bottom 10% of collected probabilities. After the binary classification, short segments under 200 ms were omitted from output.

4. EVALUATION

Firstly, we evaluate the difference in classification performance between homoogeneous recognition scenario and heterogeneous recognition scenario. Secondly, we try to find the best feature normalization method and data scope in heterogeneous recognition scenario.

4.1. Setup

To evaluate the classification performances in the homogeneous recognition scenario, we perform a 4-fold cross validation on the TUT-vocal-2016 dataset. The evaluation results are reported averaging the four folds.

Scenario	Normalization data scope		MVN	QE
	Training	Testing		
Heterogeneous	Global	Global	69.6	
Homogeneous	Global	Global	95.5	
Heterogeneous	Recording	Recording	72.7	76.2
	Dataset	Dataset	88.1	91.6
	Subdataset	Subdataset	96.8	93.9
	Subdataset	Dataset	90.7	90.4
	Dataset	Recording	81.1	78.3
	Subdataset	Recording	81.2	81.1

Table 2: Evaluation on different data scopes using mean-variance feature normalization (MVN) and quantile equalization (QE).

In the evaluation of the heterogeneous recognition scenario, TUT-vocal-2016 dataset is used for testing, while the rest of the datasets are used for training. The baseline is global feature normalization, where all the feature vectors in training and testing material are operated with the same linear transformation based on the statistics of training material alone. Two feature normalization techniques, MVN and QE are evaluated, along with three feature normalization data scopes, recording-wise, dataset-wise, subdataset-wise. Particularly, we evaluate recording-wise normalization for the testing data, while using all the three normalization data scopes for training data. In many practical cases, the data source is unknown at the recognition stage, or the statistics of the whole recognition dataset are not available.

4.2. Results

The experimental results are reported in unweighted accuracy (average recall), of the two classes. The experimental results are shown in Table 2. MVN and QE give similar performance through all the experiments. In contrast, the feature normalization data scope significantly affects the classification performances. Based on that, we can simply use the results from MVN to discuss different normalization data scopes.

The obtained accuracy using subdataset-wise normalzation was remarkably high. We investigated the clustering results of the TUT-vocal-2016 dataset and found that the speech and singing recordings were clustered to different subdatasets. All of our training datasets consist either speech or singing. The condition is more matched, when training and testing data scope contains only just one class, which leads to a big improvement when the testing dataset is normalized subdataset-wise. When online application is considered (normalization scope is recording-wise at recognition), there is no major difference in performance between dataset-wise and subdataset-wise normalization.

In most robust speech recognition studies [6, 8], QE gives clearly better performance than MVN. However, this conclusion does not hold in our study. In the robust speech recognition studies, the purpose of feature normalization is to improve the noise robustness. In comparison, all the datasets used in our study are recorded in close microphone scenario, thus relatively clean from interfering sounds. Our experimental results suggests that it has no benefit using QE compared to MVN, when the mismatch is mainly on channel effect.

5. CONCLUSION

This paper targets on a generalized vocal mode classifier, which is able to perform classification on signals from arbitrary data sources. A new dataset TUT-vocal-2016, containing both speech and singing from 20 volunteers, was collected for evaluation.

In a homogeneous recognition scenario, a four fold cross-validation is made on TUT-vocal-2016 alone. In a heterogeneous recognition scenario, four speech datasets and three singing datasets are used as training material, and TUT-vocal-2016 is used for testing. In the experiments, the vocal mode classifiers were based on log-Mel band energies as features and multi-layer perceptrons as models. The experimental results showed that the classifier gave clearly higher accuracy 95.5% in the homogeneous recognition scenario compared to heterogeneous recognition scenario (69.6%).

This result shows that the classification performance is severely degraded by training-recognition mismatch. However, we found no public evaluation setup for sound classification targeting on heterogeneous recognition scenario. A new evaluation setup should be established to test the capability of classifiers to work on heterogeneous data sources.

Various feature normalization methods were tested to improve the classification performances in the heterogeneous recognition scenario. Subdataset-wise mean-variance normalization was found to gave the best performance, which achieved a classification accuracy of 96.8%. However, the subdataset-wise normalization relies on the knowledge of recognition data source and a sufficient amount of data from the source is needed to estimate the feature distribution. In case that the feature distribution can only be estimated based on the current signal to be recognized, the best achieved accuracy was 81.2 %.

This suggested that an online application would be much more challenging than an offline application for a heterogeneous recognition scenario. In the future, studies should be made on normalization methods that requires less data to improve on heterogeneous recognition scenario.

6. REFERENCES

[1] T. Drugman, Y. Stylianou, Y. Kida, and M. Akamine, "Voice activity detection: Merging source and filter-based information," *IEEE Signal Process. Lett.*, vol. 23, no. 2, pp. 252–256, 2016. [Online]. Available: http://dx.doi.org/10.1109/LSP.2015.2495219

[2] F. G. Germain, D. L. Sun, and G. J. Mysore, "Speaker and noise independent voice activity detection," in *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013*, 2013, pp. 732–736. [Online]. Available: http://www.isca-speech.org/archive/interspeech_2013/i13_0732.html

[3] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proceedings of the ACM International Conference on Multimedia*, 2014, pp. 1041–1044. [Online]. Available: http://doi.acm.org/10.1145/2647868.2655045

[4] K. J. Piczak, "ESC: dataset for environmental sound classification," in *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference, MM '15*, 2015, pp.

1015–1018. [Online]. Available: http://doi.acm.org/10.1145/2733373.2806390

[5] T. Virtanen, A. Mesaros, T. Heittola, M. Plumbley, P. Foster, E. Benetos, and M. Lagrange, *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*. Tampere University of Technology. Department of Signal Processing, 2016.

[6] S. Molau, F. Hilger, and H. Ney, "Feature space normalization in adverse acoustic conditions," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '03*, 2003, pp. 656–659. [Online]. Available: http://dx.doi.org/10.1109/ICASSP.2003.1198866

[7] F. Hilger and H. Ney, "Quantile based histogram equalization for noise robust large vocabulary speech recognition," *IEEE Trans. Audio, Speech & Language Processing*, vol. 14, no. 3, pp. 845–854, 2006. [Online]. Available: http://dx.doi.org/10.1109/TSA.2005.857792

[8] Á. de la Torre, A. M. Peinado, J. C. Segura, J. L. Pérez-Córdoba, M. C. Benítez, and A. J. Rubio, "Histogram equalization of speech representation for robust speech recognition," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 3, pp. 355–366, 2005. [Online]. Available: http://dx.doi.org/10.1109/TSA.2005.845805

[9] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication*, vol. 25, no. 1-3, pp. 133–147, 1998. [Online]. Available: http://dx.doi.org/10.1016/S0167-6393(98)00033-8

[10] D. Pearce and H. Hirsch, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Sixth International Conference on Spoken Language Processing, ICSLP 2000 / INTERSPEECH*, 2000, pp. 29–32. [Online]. Available: http://www.isca-speech.org/archive/icslp_2000/i00_4029.html

[11] E. Cakir, T. Heittola, H. Huttunen, and T. Virtanen, "Polyphonic sound event detection using multi label deep neural networks," in *2015 International Joint Conference on Neural Networks, IJCNN*, 2015, pp. 1–7.

[12] J. Li, W. Dai, F. Metze, S. Qu, and S. Das, "Comparison of deep learning methods for environmental sound detection," in *2017 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '17*, 2017, pp. 126–130.

[13] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. D. Green, "The PASCAL chime speech separation and recognition challenge," *Computer Speech & Language*, vol. 27, no. 3, pp. 621–633, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.csl.2012.10.004

[14] J. Kominek, A. W. Black, and V. Ver, "CMU arctic databases for speech synthesis," Carnegie Melon University, Tech. Rep., 2003.

[15] F. Hilger, S. Molau, and H. Ney, "Quantile based histogram equalization for online applications," in *7th International Conference on Spoken Language Processing, ICSLP2002*, 2002. [Online]. Available: http://www.isca-speech.org/archive/icslp_2002/i02_0237.html

[16] S. Molau, F. Hilger, D. Keysers, and H. Ney, "Enhanced histogram normalization in the acoustic feature space," in *7th International Conference on Spoken Language Processing, ICSLP2002*, 2002. [Online]. Available: http://www.isca-speech.org/archive/icslp_2002/i02_1421.html

[17] R. Hariharan and O. Viikki, "On combining vocal tract length normalisation and speaker adaptation for noise robust speech recognition," in *Sixth European Conference on Speech Communication and Technology, EUROSPEECH 1999*, 1999. [Online]. Available: http://www.isca-speech.org/archive/eurospeech_1999/e99_0215.html

[18] S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2nd ed. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1998.

[19] F. Chollet, "keras," https://github.com/fchollet/keras, 2015.

[20] Theano Development Team, "Theano: A Python framework for fast computation of mathematical expressions," *arXiv e-prints*, vol. abs/1605.02688, May 2016. [Online]. Available: http://arxiv.org/abs/1605.02688

[21] M. Grimaldi and F. Cummins, "Speaker identification using instantaneous frequencies," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 16, no. 6, pp. 1097–1111, 2008.

[22] "Karaokeversion," www.karaoke-version.com, accessed: 11.03.2016.

[23] R. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. Bello, "MedleyDB: A multitrack dataset for annotation-intensive mir research," in *15th International Society for Music Information Retrieval Conference*, 2014.

[24] J. Mora, F. Gomez, E. Gomez, F. Escobar-Borrego, and M. Diaz-Banez, "Melodic characterization and similarity in a cappella flamenco cantes," in *11th International Society for Music Information Retrieval Conference*, 2010.

[25] T. Giannakopoulo, "pyaudioanalysis: An open-source python library for audio signal analysis," *PloS ONE*, 2015.