sciendo

NP13

Creak Rate Variation in Individual Speakers of Finnish

Michael L. O'Dell[1], *Tommi Nieminen*[1,2], *Liisa Mustanoja*[2]

[1]University of Helsinki
[2]Tampere University
michael.odell@helsinki.fi, tommi.k.nieminen@tuni.fi, liisa.mustanoja@tuni.fi

ABSTRACT

We report on an investigation into the use of creaky voice in Finnish speech using a corpus of individuals recorded three times in their lives over a span of approximately 40 years. Classification of creak by hand is very time consuming. Automatic detection of creak on the other hand is unreliable, especially with non-optimal field recordings, such as the ones in our corpus. Using sparse hand marked data to train a classification algorithm applied to large scale automatic data, we have been able to combine evidence from both sources and extend the scope of the analysis and increase the precision of estimates. Our results indicate that some individuals living in Helsinki and Tampere have changed the amount of creak they use during their lifetime. However, no general trend was observed for all speakers examined. All of our speakers used at least some creak already in the 1970s, some very little and others quite a lot. Use of creak can be fairly stable for an entire one hour interview, but it can also be fairly volatile, varying within a single interview. In addition to the proportion of creak used by speakers, we have also begun to investigate where creak is more likely to occur. Results indicate that creak is more prevalent in the final half of interpause intervals, which gives some support to suggestions that creak may be associated with end of utterance in Finnish or that creak may be one way for a speaker to conserve airflow.

Keywords: Finnish, creaky voice, longitudinal corpus

1. Introduction

In this paper we present an overview of our ongoing analysis of creak as part of a larger project analyzing a longitudinal corpus of Finnish speakers in the 1970s, 1990s and 2010s [1, 2]. We have been looking at various prosodic features including how or indeed whether individual adults' $F0$ distributions have changed over this period of approximately 40 years [3]. We became interested in the incidence of creaky voice in our corpus, originally because of the difficulties creaky voice causes for $F0$ measurement, but also to see whether individuals' use of creaky voice has changed over the years.

Languages always change, speech communities are varied, and idiolects also change in many ways. We know from previous work with this corpus that $F0$ distribution can change dramatically during the course of adulthood [3]. For instance, we found that women's $F0$ dropped between the ages 20 and 40, regardless of when they were born. Also, the range of $F0$ (dispersion) increased for almost all subjects from age 20 to 40 to 60.

How much can phonation behavior, such as use of creak, change in the lifetime of an individual? It has been suggested that use of creak is increasing not just in English [4], but in Finnish as well [5, 6]. Is such a trend observable in the behavior of individual Finnish speakers from 1970 to 1990 to 2010?

Table 1: Longitudinal corpus of Finnish spoken in Helsinki [1]; Longitudinal corpus of Finnish spoken in Tampere [2]

Helsinki	Tampere	
1970	1977	interviews with informants in 3 age groups: approximately 20-, 40- and 60-year olds
1990	1997	former informants interviewed again (Helsinki & Tampere), additional young informants recruited (Helsinki)
2010	2019	informants interviewed a 3rd time, informants interviewed a 2nd time, additional young informants

2. Corpus

The approximately hour long interviews utilized in this research belong to a longitudinal corpus of speakers living in two major Finnish cities, Helsinki and Tampere, collected for use in sociolinguistic research starting in the 1970s ([1, 2], cf. Table 1). During these interviews the interviewers speak very little, most of the speech is pro-

duced by the interviewee. The corpus represents a combination of panel survey (some of the same speakers were reinterviewed at successive periods) and trend survey (some of the speakers were chosen based on matching background variables) [cf. 7, pp. 43–112].

Six speaker subcorpus. In this article we restrict attention to six speakers from the panel survey part of the corpus, that is, speakers that were recorded three times, in the 1970s, 1990s and 2010s. Such a panel survey comprising three time points is still relatively rare world wide. Three of our speakers are from Helsinki and three are from Tampere. All of them belong to the same age group (or cohort), the only one that was interviewed three times, approximately 20 years old at the time of the first interview in the 1970s. In what follows they are referred to using the pseudonyms *Anneli*, *Anita*, *Antti* (from Helsinki), *Taina*, *Tiina*, *Tuomo* (from Tampere). Our subcorpus thus includes a total of 18 interviews. In the future we plan to examine more speakers and other age cohorts from this corpus.

Hand marked phonation data. For the six speakers in our subcorpus we have available a small amount of hand marked labels classifying speech into four phonation based categories: **U** = unvoiced (e.g. stops, fricatives), **V** = chest voice, **C** = creak, and **W** = whisper. Because marking by hand is very time consuming, we only have a sparse random sample of hand classified speech for each interview.

Sampling was carried out as follows: Random measurement points were generated using a script in Praat [8], and the speech signal (excluding pause) within an approximate 1.2 s window surrounding the measurement points was classified by hand into intervals marked **U**, **V**, **C**, or **W**. Sampling continued until at least 200 intervals were labeled covering a total of at least 30 s of speech sampled from the first quarter of each interview.

Annotation was carried out by a professional phonetician (the first author) in Praat using both the spectrogram display and the signal display and then adjusting boundaries by checking intervals auditorily with headphones. The acoustic criteria utilized were those enumerated in [9], although no attempt was made to distinguish between different types of creak.

3. Analysis of hand marked data

The raw percentages for our sparse hand marked data are shown in Fig. 1. It would appear that the amount of creak may differ from speaker to speaker and year to year, but to make inferences based on this data we need to use a statistical model. Estimating uncertainty is necessary for judging any differences we may be interested in—e.g. "How sure are we that percentage of creak has increased?" With a suitable statistical model Bayesian inference can be based on posterior distributions of the parameters in the model.

In the present case, it is important to model durations (or putting it another way, the density of shifts in phonation), not just raw percentages, in order to assess uncertainty. For instance, suppose for the sake of argument that periods of creaky voice (and non-creaky voice) lasted on average for several hours. In that case a one hour sample would tell us relatively little. If, on the other hand they alternated several times each minute, a one hour sample would yield a fairly precise estimate of creak percentage.

One possible statistical model is a simple Markov process in which there is a set of probabilities (dependent on the present state) that the process will stay in its present state or jump to any other possible state. In the present case the states would be the possible phonation classes, with pause acting as starting and stopping states. A Markov process implies that so-called sojourn times (the time the process remains in one state) follow an exponential distribution. This in turn means that the standard deviation of log sojourn time is $\pi/\sqrt{6} \approx 1.28255$. For our hand marked data the sample standard deviation of log durations for phonation states ranges from about 0.25 to 0.95. This clearly indicates that a simple Markov model is inadequate for our data.

Instead of a simple Markov model, we employ a semi-Markov process model, which includes separate Log-Normal duration (sojourn) distributions. This model is illustrated in Fig. 2. Each possible state has its own Log-Normal distribution (with two parameters, mean and variance) indicating how long the process remains in that state. At the end of this duration the process changes state according to a set of probabilities (dependent on current state), just as in the simple Markov model (except that transition to the same state is not allowed). Naturally, all parameters (transition prabability, means and variances for duration distributions) are allowed to vary by speaker and interview.

Using two parameter Log-Normal distributions for duration allows more flexibility than the simple Markov model described above (which is equivalent to a semi-Markov model with single parameter exponential duration distributions). For still more flexibility three parameter Weibull distributions have also been used [10].

Posterior distributions for standard deviations of effects in the semi-Markov model are shown schematically in Fig. 3. These give an indication of how important the various effects are. If an effect is completely irrelevant, its standard deviation would be zero, since all parameters differing only in that effect should be the same. A large standard deviation indicates that parameters for the effect in question differ a great deal and the effect is therefore relatively important [11, 12].

It can be seen in Fig. 3 that the magnitude (or mean) parameter and dispersion (or standard deviation) parameter both vary greatly according to phonation type, which is not unexpected. For instance, overall, mean duration of voiced segments is greater than the mean duration of other types. Any overall (main) effect of interview year on duration distibution is relatively small, as is the main effect for speaker. Most interactions for duration distribution parameters are also fairly small, with the exception of the TYPE×SPEAKER interaction for duration distribution mean. This interaction indicates that the differences in magnitude of durations for different types (for instance, how much longer voiced segments are on average compared to creaky segments) are also different for different speakers. In other words, some speakers spend a longer time in a certain state than other speakers do. There are also sizeable differences between speakers in the transition probability parameters, as indicated by the SPEAKER effect for transition probabilities in Fig. 3 (YEAR and YEAR×SPEAKER effects are also definitely non-zero, but smaller in size).

While these are interesting details, the main purpose of using this model is to be able to estimate the amount or proportion of creak being used. With the semi-Markov model it is possible to calculate the expected proportion of time spent in each state, and we explore these main results in Section 5 for the case of creak (see also Fig. 7), after first considering the use of automatic creak detection.

4. Utilizing automatic creak detection algorithms

Because hand marking is very time consuming, it would be advantageous to be able to detect creak in the audio signal automatically. However, automatic classification of creak is a very challenging task, especially with non-optimal field recordings, such as the ones in our corpus. So far attempts to find an ade-

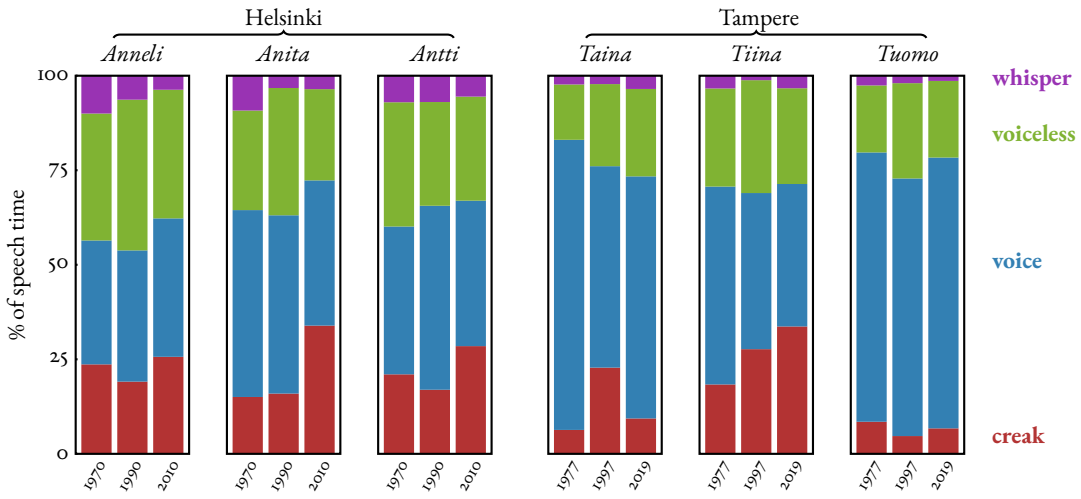

Figure 1: Sparse hand classified data for six speakers

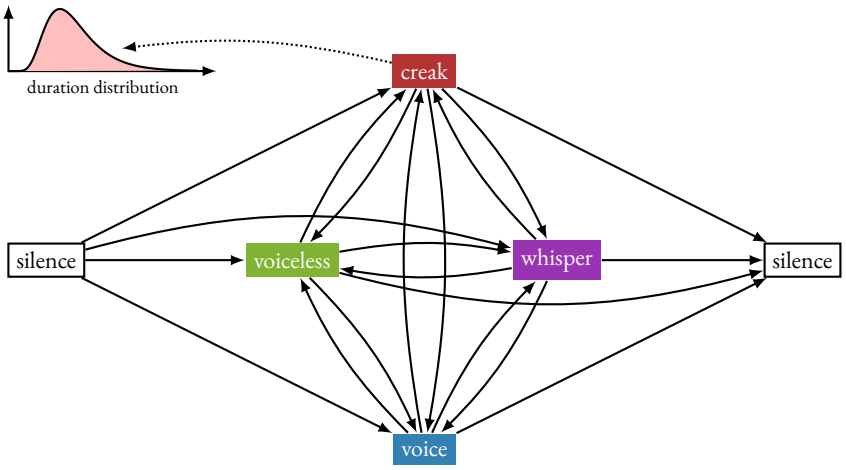

Figure 2: Semi-Markov process model

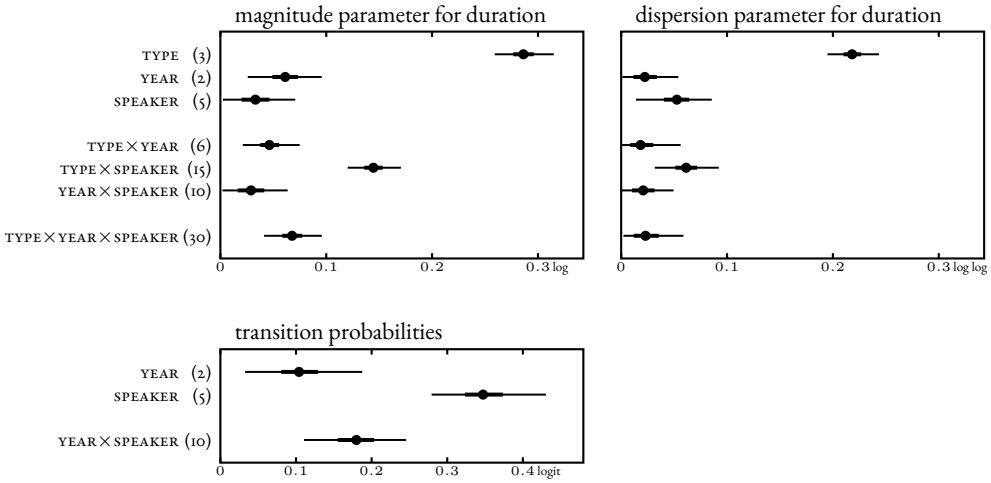

Figure 3: Estimated importance (standard deviations) of effects in the semi-Markov model: posterior median, 50 % CI, 95 % CI; degrees of freedom in parentheses

quate creak detection algorithm to use with our recordings have been disappointing. In what follows we utilize two algorithms: `detect_creaky_voice()` in the COVAREP-repository ([13], [14], hereafter referred to as COVAREP) and `creakByRoughness()` based on the concept of psychoacoustic roughness ([15], [16], hereafter referred to as ROUGH). Both are available as MATLAB software packages [17]. Because initial investigation revealed that signal intensity can have a large effect on the outcome of these algorithms, all recordings were normalized to 70 dB using Praat [8] prior to analysis.

4.1. Evaluation of algorithms and comparison with hand marked data

COVAREP correlates with our tiny hand marked samples fairly well for some interviews, but fairly poorly for others. The so-called Area Under the Receiver Operating Curve (AUROC or simply A') provides one way to measure this. It is a measure of discriminating power, which can be interpreted as the probability that a randomly selected creaky frame gives a higher COVAREP value than a randomly selected non-creaky frame, so that $A' = 0.5$ would indicate no separation of creak and non-creak based on COVAREP, whereas $A' = 1$ would indicate complete separation [18]. Estimated A' for COVAREP applied to our hand marked data ranges from 0.768 to 0.977 (see Table 2).

Table 2: AUROC (A'): Estimated probability that a randomly selected creaky sample receives a higher value than a randomly selected non-creaky sample

		1970	1990	2010
Anneli	COVAREP	0.928	0.945	0.834
	ROUGH	0.825	0.794	0.680
Anita	COVAREP	0.936	0.915	0.862
	ROUGH	0.737	0.653	0.665
Antti	COVAREP	0.922	0.768	0.790
	ROUGH	0.707	0.665	0.650
		1977	1997	2019
Taina	COVAREP	0.866	0.920	0.945
	ROUGH	0.682	0.751	0.820
Tiina	COVAREP	0.865	0.905	0.933
	ROUGH	0.708	0.688	0.752
Tuomo	COVAREP	0.871	0.808	0.977
	ROUGH	0.561	0.673	0.667

Possible COVAREP scores fall in the range $[0, 1]$ and are meant to be interpreted as probability of creak, but they do not generally correspond to estimates of creak percentage based on the hand marked sample. In fact the level and reliability of the COVAREP scores can vary drastically for the same speaker from year to year, so it is not possible to determine a single threshold for all interviews—often any inference regarding increase or decrease of creak would be possible depending on the choice of threshold.

Similar comments apply to the ROUGH scores, although in general ROUGH fares a bit worse than COVAREP. Estimated A' for ROUGH applied to our hand marked data ranges from 0.561 to 0.825 (see Table 2). Possible ROUGH scores fall in the range $[0, \infty)$, and can be easily converted to a supposed creak probability with a score of one corresponding to $p = 0.5$, i.e. to the boundary between creak being less probable or more probable than not. Again, these raw scores do not generally correspond to estimates of creak percentage based on the hand marked sample, and the level and reliability of the ROUGH scores can vary drastically even for the same speaker from year to year, just as was the case for the COVAREP scores.

4.2. Combining evidence

Ideally we would like to combine the evidence from our reliable but sparse hand classification with large scale automatic classification data, the reliability of which is open to question. The situation is shown schematically in Fig. 4. At some points in time, for instance point (a) in Fig. 4, we have information from three sources. Most of the time, however, for instance point (b) in Fig. 4, we only have the COVAREP and ROUGH scores available.

In principle, expanding the statistical model to take into account the additional information provided by COVAREP and ROUGH scores is straight forward from the point of view of Bayesian analysis. We can simply model the additional scores as coming from distributions depending on the sometimes known (hand classified), but mostly unknown states. Computationally, however, this is challenging. The hand marked data used for the semi-Markov model consists of durations of observed continuous stretches of creak and other states, combined with observed transitions from one state to another. Automatic detection data, on the other hand, is sampled (typically every 10 ms), so we need to convert to a *discrete* (ie. sampled) version of the model, leading to an explosion of data, even for the relatively small samples of speech corresponding to the hand classified data. This drawback is made worse by the fact that the goal is to utilize entire interviews each lasting approximately an hour, so that computation virtually comes to a standstill. Optimizing computational routines can help somewhat, but not enough to make the use of an expanded semi-Markov model feasible.

Of course it is an open question whether adding large amounts of unreliable data will actually increase the precision and reliability of estimates. If the computational challenge could be tackled, we would be able to rapidly expand analysis to our entire database, which includes many more speakers and different age groups.

4.3. Probabilistic classification trained with hand marked data

Given a probability p_i of classification as creak for each sample i, a rough estimation of credible intervals (CI) for total percentage of creak can be estimated using the Poisson-Binomial distribution (remembering to divide by the sample size to estimate a proportion rather than the total number of creaky samples). Exact quantiles for Poisson-Binomial distributions can be computed [19], [20], but even this is computationally prohibitive when the number of samples is high as in the present case. An easily computed alternative is to use simulation to calculate quantiles. This has the added advantage that uncertainty in the individual creak probabilities p_i can be handled at the same time.

The value of the COVAREP creak function is already in the form of a probability ($0 \leq p_i \leq 1$), so in principle this procedure could be applied to the raw COVAREP scores. However, it is clear that these raw scores interpreted as probabilities do not correspond to creak percentages, at least for the hand marked data set (most often they greatly underestimate the amount of creak). Likewise the raw roughness scores ($0 \leq r_i < \infty$) can easily be converted to probabilities (using $p_i = r_i/(1 + r_i)$), but again these values do not provide reliable estimates of creak percentage.

An alternative procedure is to train a probabilistic classification algorithm as a function of the raw scores. There are many types of such algorithms, here we utilize a simple (easily calculated and easy to interpret) logistic regression model based on the raw scores trans-

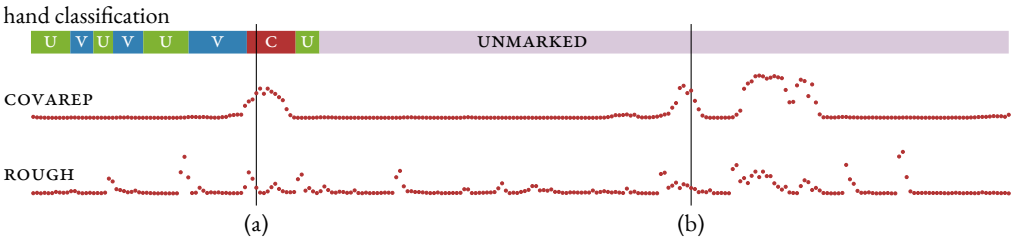

Figure 4: Combining evidence from three sources

formed to the real plane.

$$(c_i, r_i) \to (x_i, y_i) = (\text{logit}(c_i), \log(r_i)) \quad (1)$$

Estimated probability of creak can thus be expressed according to the logistic regression model as

$$\text{logit}(p_i) = \beta_0 + \beta_{\text{COVAREP}} \cdot x_i + \beta_{\text{ROUGH}} \cdot y_i \quad (2)$$

The parameters of the model ($\beta_0, \beta_{\text{COVAREP}}, \beta_{\text{ROUGH}}$) are allowed to vary by interview (17 degrees of freedom) based on the hand marked data, thus taking into account the clear differences in how the raw scores behave for different recordings.

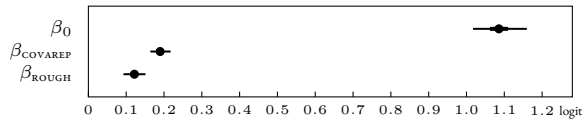

Figure 5: Estimated standard deviations of logistic regression parameters; 17 degrees of freedom

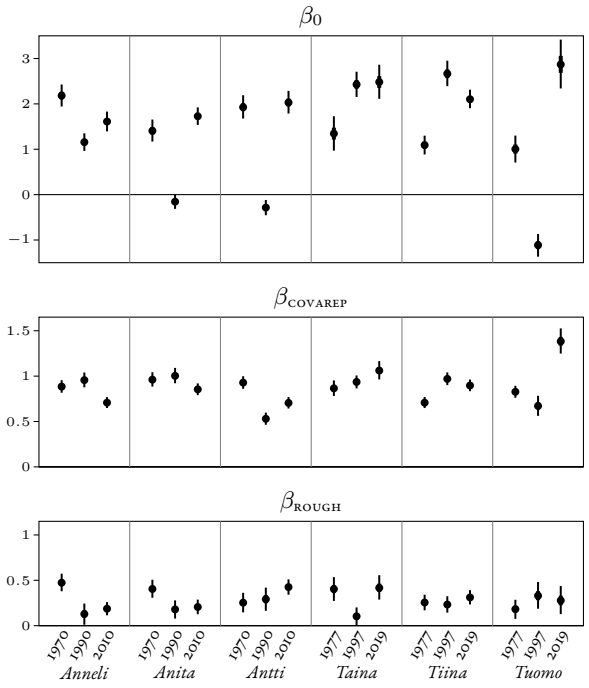

Figure 6: Estimated logistic regression parameters for all interviews

The results of fitting a Bayesian logistic classification model to our (hand marked) data are shown in Fig. 5, showing estimated standard deviations of parameters, and Fig. 6, showing estimates of individual parameters. From the posterior standard deviation estimates (Fig. 5), it can be seen that the intercept parameter β_0 varies widely (has a large, non-zero standard deviation), indicating that different interviews vary widely in how much the automatic measures underestimate (or overestimate) creak probability. If $\beta_0 = 0$, then the raw score $(c_i, r_i) = (0.5, 1)$ corresponds to the expected 50 % creak boundary; if $\beta_0 > 0$, then creak is under-estimated (creak probability is higher than the raw scores indicate), and if $\beta_0 < 0$, then creak is over-estimated. It can be seen in Fig. 6 that creak is generally underestimated, with the exception of *Anita* : 1990, *Antti* : 1990 and especially *Tuomo* : 1997, for which creak is overestimated.

Posterior standard deviation estimates for β_{COVAREP} and β_{ROUGH}, are much smaller, but are also clearly non-zero (cf. Fig. 5), indicating that the interviews also differ in how clearly the scores separate creak from non-creak. Estimates of β_{COVAREP} and β_{ROUGH} are all positive (Fig. 6), which means that larger scores do correspond to greater creak probability, as expected (negative values would mean the opposite). The larger these parameters are, the more effectively the scores separate creak from non-creak. The fitted logistic regression model provides the required estimated creak probability for each sample, given the values of COVAREP and ROUGH, so that based on these probability estimates a creak percentage CI can be calculated, for an entire interview or for some a priori defined subset, taking into account the uncertainty of the estimates by using simulation as discussed above.

The final result is not a fully Bayesian analysis, since COVAREP and ROUGH scores are taken as given, rather than modeled as the result of a stochastic process. Also, unlike the semi-Markov model applied above to the hand marked data, the model ignores the dynamic aspect of creak, even though creak probabilities at nearby times are obviously not independent. However, this approximate model does not assume *absolute* independence, but only *conditional* independence of samples (given the explanatory variables COVAREP and ROUGH), so this is perhaps not so problematic, and we may hope that much or even most of the interdependence between nearby samples is captured by the COVAREP and ROUGH scores.

5. Results and discussion

We now turn to a discussion of various questions considered in the light of the available data and the results of the statistical analyses.

Fig. 7 shows several estimates of the percentage of creaky voice for each of our six speakers' three interviews. First of all, for convenience, the raw percent of creak calculated from the sparse hand marked data is copied from Fig. 1. Also shown is an estimate (posterior distribution characterized by median, 50 % CI and 95 % CI) of the expected creak percent based on a Bayesian analysis of the

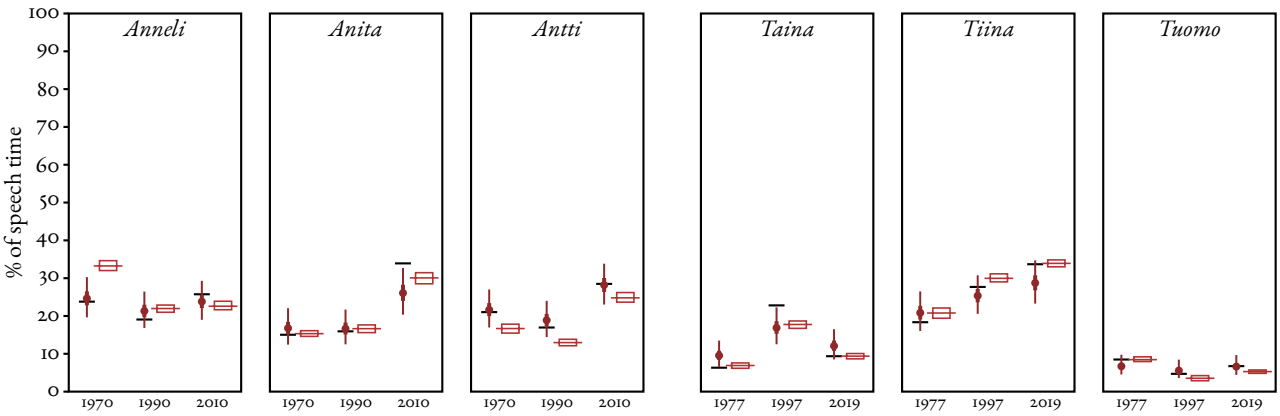

Figure 7: Various estimates of creak percent. —: raw percent for hand marked data; ⧫: posterior median, 50 % CI, 95 % CI for semi-Markov model applied to hand marked data; ⊟: estimated median and 95 % CI for entire interview, based on COVAREP and ROUGH logistic regression model trained with hand marked samples.

semi-Markov model applied to the sparse hand marked data. This is the amount of creaky voice expected in the long run, given that the stochastic process characterized by the semi-Markov model continues unchanged. The posterior distribution indicates uncertainty left after incorporating the data, it does not indicate expected variation in the process itself. Lastly, an estimated median and 95 % CI is shown for the portion of creaky voice in each interview, based on the hybrid analysis using a trained logistic regression model applied to the COVAREP and ROUGH scores sampled every 10 ms for the entire interview.

There are some obvious differences in the estimates for the different anaylses. First of all, the 95 % CI are much narrower for the logistic regression analysis compared to the semi-Markov analysis. This is perhaps expected, given that it is based on much more data. It does, in any case, show that adding a large amount of relatively unreliable data did result in more precise estimates. In most cases the two analyses show similar patterns, which also gives us confidence in the results. There are also some interesting discrepancies, which we discuss further in Section 5.2.

5.1. Has creak increased in Finnish?

It has been suggested that use of creak is increasing in some languages, at least in English [4], but also including Finnish [5]. Can we see evidence of this in our data?

Some changes between interviews in the amount of creak an individual uses are apparent already in the less precise estimates from the semi-Markov analysis. It is fairly clear that creak percentage has *increased* for *Anita* and *Antti* from 1990 to 2010, and for *Taina* from 1977 to 1997. These differences are born out by the more precise estimates from the logistic regression analysis. Also additional clear differences emerge due the increased precision, the largest being a *decrease* in creak for *Anneli* from 1970 to 1990 (although she still has a large amount of creak in 1990 and 2010), a *decrease* for *Taina* from 1997 to 2019, and an *increase* for *Tiina* from 1977 to 1997 and from 1997 to 2019.

These major changes are summarized schematically in Table 3. It can be readily observed that if there is a possible weak overall rising trend, certainly not all of our speakers follow this trend. There are many changes in level of creak to be seen within the lifetime of individuals, but with only six speakers we already see at least five different patterns of change over the three interviews.

Table 3: Change in creak percent for individuals

	1970s → 1990s → 2010s (age 20 → age 40 → age 60)	
Anneli	↘	→
Anita, *Antti*	→	↗
Taina	↗	↘
Tiina	↗	↗
Tuomo	→	→

5.2. How stable is creak rate?

In the case of $F0$, we found previously that a speaker's distribution is fairly constant over the period of a single interview, and measurements restricted to a total of slightly more than 30 s sampled throughout the first quarter were sufficient to provide an excellent approximation of this distribution [3]. What about creak? Is a 30 s sample enough? Is an hour long sample enough? Can creak rate vary within an hour long interview as much or more than it does from one interview to the next, 20 years later?

To help answer this question we divided each interview into six sections approximately ten minutes in duration, always dividing at a pause. As mention above, one of the advantages of the logistic regression analysis is the possibility of estimating a CI for creak percentage restricted to some a priori grouping. Here the samples are grouped into successive periods of speech within each interview. The results, showing how creak level varies within the interview, are shown in Fig. 8. It is immediately obvious, at least for some speakers, that some 10 minute periods in an interview are quite different than others in terms of creak rate. This is perhaps clearest for *Anneli* : 1970. The rate of creak estimated for the fifth 10 minute stretch of speech is much higher than the rest. She utilized much more creak for an extended period of time at that point in her interview. This also explains why *Anneli* : 1970 exhibits one of the largest discrepancies for the two methods of estimating creak rate, as shown in Fig. 7: the semi-Markov analysis is based on data from the first quarter of each interview, whereas the logistic regression analysis use data from the entire interview.

At the same time there does not appear to be any consistent trend (rising or falling) for running creak rate. Depending on the interview, creak may stay relatively stable throughout, it may rise, fall,

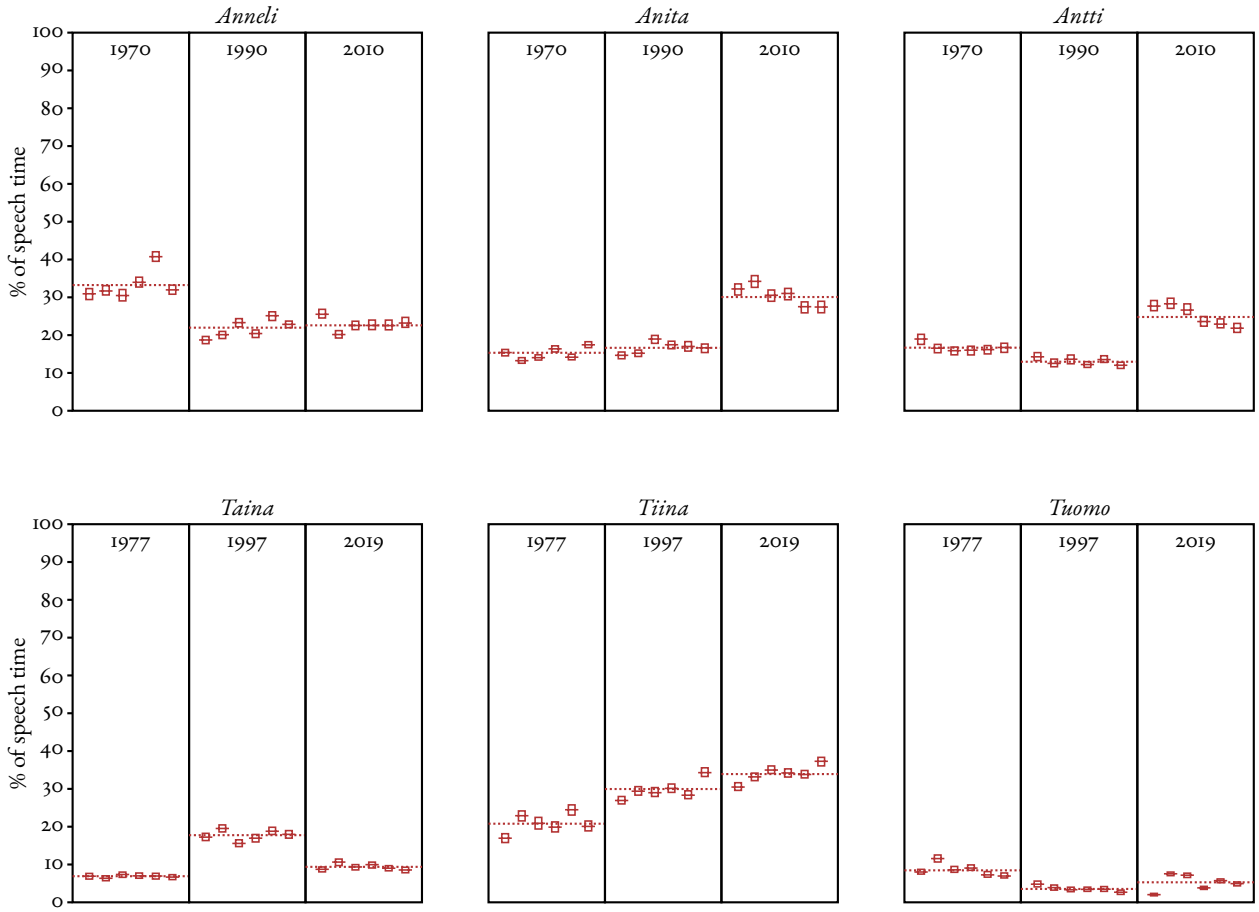

Figure 8: Estimates of running creak percent. ⊟: estimated median and 95 % CI for interview divided into approximately 10 min intervals, based on COVAREP and ROUGH trained with hand marked samples. Dotted lines indicate the estimated median for the entire interview (cf. Fig. 7).

or follow a more varied pattern. Overall, it does appear that within interview variation is generally smaller than the differences observed for one speaker between interviews, so perhaps a one hour interview does provide a fairly stable overview of an individual's use of creak at that point in life, but without data from successive hours of speech, even this conclusion is very weak.

5.3. Where is creak most likely?

In addition to proportion (or overall probability) of creak used by speakers, we have also begun to investigate where creak is more likely to occur. We have looked at two hypotheses in particular, based on the findings of other studies. If creak is a potential marker of end of utterance or end of speaker's turn in Finnish [21], we would expect to find that creak is more prevalent towards the end of interpause intervals. As far as signaling end of turn this effect might be expected to be weak at best in our corpus, given that there is little turn taking (the interviewer generally talks much less than the interviewee). Based on respiratory data, it has also been suggested that creak may be one way for a speaker to conserve airflow and therefore produce longer stretches of speech before breathing [22].

To investigate this further, we divided all samples in each interview into two groups, one comprising the initial half of each interpause interval, the other comprising the final half of each interpause interval. The results of using the logistic regression analysis to estimate CIs for creak percentage restricted to these groups is shown in Fig. 9. It can be seen that overall creak percentage is consistently higher for samples in the final half of a interpause interval. This holds true for each of our 18 interviews.

This surprisingly robust finding is compatible with the idea that creak is associated in Finnish with end of utterance. However, a similar finding ("tendency to creak towards the end of turns") has been reported for Estonian dialog [23], even though creak was found not to be associated with turn-taking. And of course creaky voice cannot always signal end of utterance, given that there is often a sizeable amount of creak in the initial halves of interpause intervals. Although we do not have respiratory data to accompany our acoustic data, the idea of conserving airflow also gains some support from the pattern in Fig. 9.

Besides predicting more creak towards the end of interpause intervals, the airflow hypothesis also suggests that creak might be more prevalent during longer interpause intervals. To investigate this, the interpause intervals for each interview were divided into four approximately equal groups based on their duration, from shortest to longest, so that the total duration for each group was approximately one quarter of the total duration of all interpause intervals combined. Estimated creak percentage for these four duration groups is shown in Fig. 10. It would appear that there may be a very weak cor-

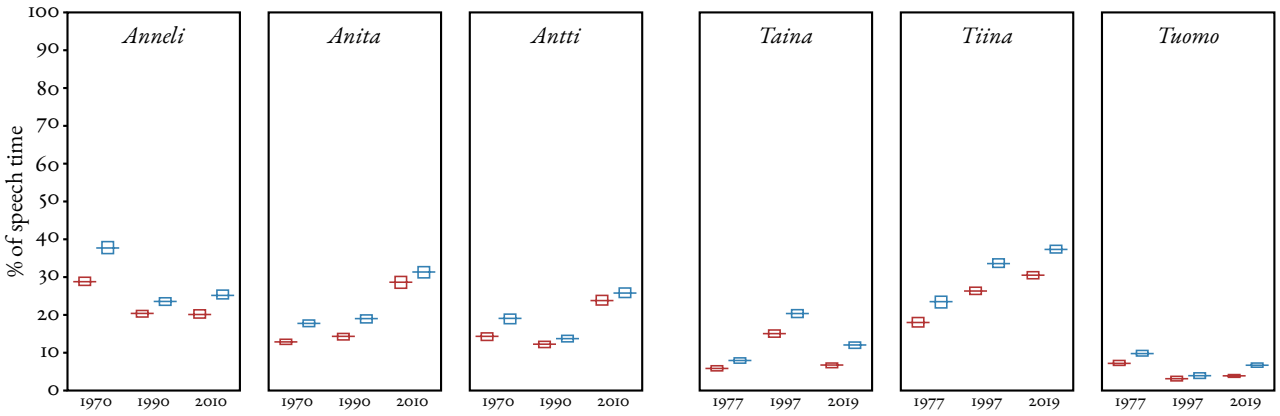

Figure 9: Estimates of creak percent within interpause intervals. ⊟: estimated median and 95 % CI for interview restricted to initial half of each interpause interval; ⊟: estimated median and 95 % CI for interview restricted to final half of each interpause interval.

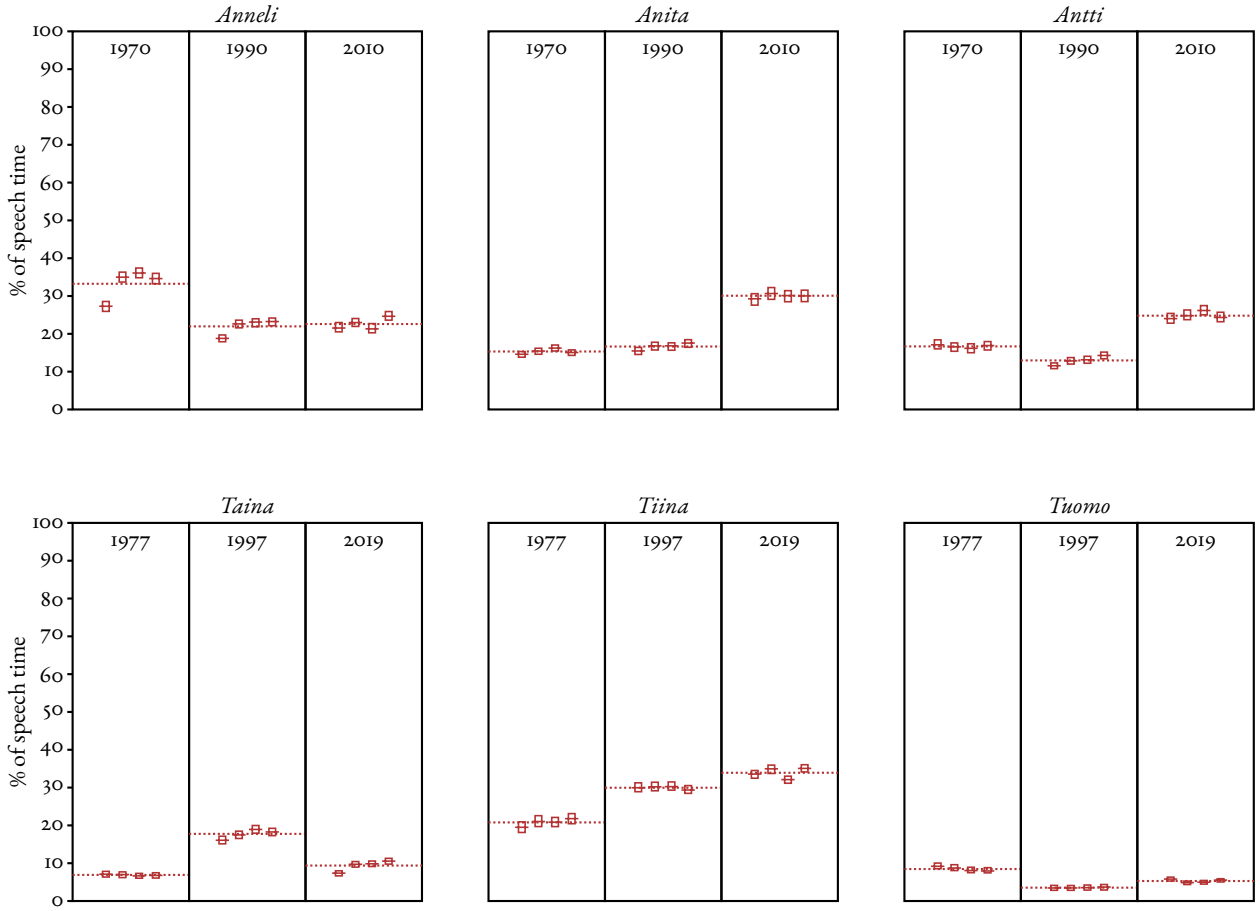

Figure 10: Estimates of creak percent by interpause interval size. ⊟: estimated median and 95 % CI for interpause intervals divided into four size groups (shortest to longest), based on COVAREP and ROUGH trained with hand marked samples. Dotted lines indicate the estimated median for the entire interview (cf. Fig. 7).

relation between creak and interpause interval duration for some interviews (clearest for *Anneli* : 1970, *Anneli* : 1990 and *Taina* : 2019), but in these cases it seems to be mostly restricted to a difference between the shortest interpause intervals (first group, with less creak) compared to the other interpause intervals.

Our findings are compatible with other (not mutually exclusive) hypotheses as well. For instance, it is known that $F0$ and intensity generally drop as speech progresses (declination), so the difference in creak percentage could reflect an association of creak with low $F0$ and low intensity [23].

6. Conclusions

6.1. Measuring creak

The availability of hand marked data is crucial to assessing creak voice behavior. The main drawback of hand marked data is the time involved in collecting it. It is possible to alleviate this drawback to some extent by collecting sparse random samples over a wide range of speech. In the present case we had sparse hand marked data covering together at least 200 phonation state intervals totaling at least 30 s, sampled from approximate 15 minutes of speech. This density of sampling is perhaps fairly adequate for reasonable estimates when creak rate is stable within the range of sampling, but it would seem that 15 minutes may not be a wide enough range, because it is possible for creak rate to vary considerably even within the approximately one hour long interviews in our corpus.

Hand marked data in the form of a random sample of observed phonation state durations and transitions between these states lends itself to efficient analysis using a dynamical stochastic process model such as the semi-Markov model considered here. Ideally we would like to investigate the effects on creak behavior of such factors as utterance length, pause length, position within the interpause interval, etc. Effects corresponding to such possible trends can easily be integrated into the dynamical statistical model for the purposes of inference, but then a much larger sample would be needed to provide sufficient data for estimation.

Automatic creak detection algorithms hold out the promise of extending analyses to large quantities of speech and many speakers. However, it is apparent that automatic classification algorithms cannot be trusted or taken at face value without checking their validity for each speech sample examined. In addition to evaluating validity, in order to be useful they need to be calibrated by comparison with reliable hand marked data. Once calibrated, however, they can be very useful and make it possible to extend and sharpen the picture available with (sparse) hand marked data alone. Of course the more reliable data is available (hand marked or otherwise), the better the calibration will be and the clearer the picture will become. Adding information from new automatic algorithms, evaluated and calibrated, can increase precision even further.

6.2. Finnish creak

Based on our research it is obvious that there is wide variation in the amount of creak Finnish speakers use, from speaker to speaker, at different stages of a speaker's life, and even at different times within a single interview.

Is creak on the rise in Finnish speech? Possibly, although speakers of Finnish have certainly used creak in the past—all of our speakers used some creak in the 1970s, four out of six used a fairly large proportion of creak (greater than 10 %). While it has been suggested that use of creak is increasing in Finnish speech [5, 6], we did not find a consistent trend for all of our speakers, although three of the six did have more creak in the 2010s than in the 1970s.

Besides differing greatly from speaker to speaker, it appears the amount of creak used can also vary during a speaker's lifetime. However, no consistent trend was found as to *how* creak varies in a lifetime. With only six speakers we see at least 5 different patterns in the direction of changes, from the 1970s to the 1990s to the 2010s, or from age 20 to 40 to 60 (cf. Table 3).

There is variation in a speaker's use of creak even within a single interview, quite sizeable in some cases. Again, however, no consistent trend was found for all interviews or all speakers. This raises the question as to whether within-speaker differences observed for interviews separated by 20 years genuinely reflect change in speakers' speech behavior, or whether they might simply reflect normal variation in speakers' daily behavior.

A surprisingly robust difference was observed for the amount of creak in the initial half of interpause intervals as opposed to the final half—for all speakers in all interviews there was more creak in the final half. This may indicate that creak is indeed a marker of utterance finality, it may be related to conserving breath, or it may simply be related to lower average $F0$ as the utterance progresses.

There is no strong evidence to link creak with duration of interpause interval, although several interviews exhibit slightly less creak for the shortest interpause intervals, and for one interview (*Anneli* : 1970) this difference is robust.

Of course, there is much more work to be done, and many open questions. We intend to continue analyzing more speakers from this large corpus. In addition to clarifying the extent of creak variation, this will allow comparison with other age groups, examination of various sociolinguistic factors, and possibly open the door to answering questions about the functionality of creak in Finnish, past and present.

7. References

1. University of Helsinki, Institute for the Languages of Finland, and Heikki Paunonen. *The Longitudinal Corpus of Finnish Spoken in Helsinki (1970s, 1990s and 2010s)*. online corpus. URN: http://urn.fi/urn:nbn:fi:lb-2014073041. 2014.

2. Tampere University, Institute for the Languages of Finland, and Liisa Mustanoja. *The Longitudinal Corpus of Finnish Spoken in Tampere (1970s, 1990s and 2010s)*. online corpus. URN: http://urn.fi/urn:nbn:fi:lb-2022090821. 2019.

3. Liisa Mustanoja, Michael O'Dell, and Hanna Lappalainen. "Helsinkiläis- ja tamperelaispuhujien äänenkorkeuden muutokset 1970-luvulta 2010-luvulle". In: *Puhe ja kieli* 42.2 (2022), pp. 121–148.

4. Katherine Dallaston and Gerard Docherty. "The Quantitative Prevalence of Creaky Voice (Vocal Fry) in Varieties of English: A Systematic Review of the Literature". In: *PLoS ONE* 15.3 (2020), e0229960 1–18.

5. Tuuli Uusitalo, Laura Nyberg, Anne-Maria Laukkanen, Teija Waaramaa, and Leena Rantala. "Has the Prevalence of Creaky Voice Increased Among Finnish University Students From the 1990's to the 2010's?" In: *Journal of Voice* (in press).

6. Anne-Maria Laukkanen and Teija Waaramaa. "Suomalaisten naisopiskelijoiden luennan perustaajuuden muutos 1990-luvulta 2010-luvulle". In: *Puhe ja kieli* 40.2 (2020), pp. 123–134.

7. William Labov. *Principles of Linguistic Change vol. 1. Internal Factors*. Language in Society 20. Oxford: Blackwell, 1994.

8. Paul Boersma and David Weenink. *Praat: Doing phonetics by computer (Version 6.2.23) [Computer program]*. Last retrieved 8 October 2022, from http://www.praat.org/. 2022.

9. Patricia Keating, Marc Garellek, and Jody Kreiman. "Acoustic Properties of Different Kinds of Creaky Voice". In: *Proceedings of the 18th International Congress of Phonetic Sciences. Glasgow, UK*. Ed. by The Scottish Consortium for ICPhS 2015. University of Glasgow. 2015. ISBN: 978-0-85261-941-4.

10. Ilenia Epifani, Lucia Ladelli, and Antonio Pievatolo. "Bayesian Estimation for a Parametric Markov Renewal Model Applied to Seismic Data". In: *Electronic Journal of Statistics* 8 (2014), pp. 2264–2295.

11. Andrew Gelman. "Analysis of Variance—Why it is More Important Than Ever (with discussion)". In: *The Annals of Statistics* 33.1 (2005), pp. 1–53. URL: http : / / www . stat . columbia.edu/~gelman/research/published/AOS259.pdf.

12. Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2007.

13. John Kane, Thomas Drugman, and Christer Gobl. "Improved Automatic Detection of Creak". In: *Computer Speech & Language* 27.4 (2013), pp. 1028–1047.

14. Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. "COVAREP—A Collaborative Voice Analysis Repository for Speech Technologies". In: *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy 2014*. 2014.

15. Julián Villegas, Konstantin Markov, Jeremy Perkins, and Seunghun J. Lee. "Prediction of Creaky Speech by Recurrent Neural Networks Using Psychoacoustic Roughness". In: *IEEE Journal of Selected Topics in Signal Processing* 14.2 (2020), pp. 355–366.

16. Julián Villegas, Jeremy Perkins, and Seunghun Lee. "Psychoacoustic Roughness as Creaky Voice Predictor". In: *The Journal of the Acoustical Society of America* 140.4 (2016), pp. 3394–3394.

17. *MATLAB version 9.13.0.2049777 (R2022b)*. The Mathworks, Inc. Natick, Massachusetts, 2022.

18. James Fogarty, Ryan S. Baker, and Scott E. Hudson. "Case Studies in the use of ROC Curve Analysis for Sensor-Based Estimates in Human Computer Interaction". In: *GI '05: Proceedings of Graphics Interface 2005*. Canadian Human-Computer Communications Society, 2005, pp. 129–136.

19. Yili Hong. "On Computing the Distribution Function for the Poisson Binomial Distribution". In: *Computational Statistics and Data Analysis* 59 (2013), pp. 41–51.

20. William Biscarri, Sihai Dave Zhao, and Robert J. Brunner. "A Simple and Fast Method for Computing the Poisson Binomial Distribution Function". In: *Computational Statistics and Data Analysis* 122 (2018), pp. 92–100.

21. Richard Ogden. "Turn Transition, Creak and Glottal Stop in Finnish Talk-in-interaction". In: *Journal of the International Phonetic Association* 31.1 (2001), pp. 139–152.

22. Kätlin Aare, Pärtel Lippus, Marcin Włodarczak, and Mattias Heldner. "Creak in the Respiratory Cycle". In: *Proc. Interspeech 2018*. 2018, pp. 1408–1412. DOI: 10.21437/Interspeech.2018-2165.

23. Kätlin Aare, Pärtel Lippus, and Juraj Šimko. "Creaky Voice in Spontaneous Spoken Estonian". In: *XXVIII Fonetiikan päivät — Turku 25.–26. lokakuuta 2013*. Ed. by Katri Jähi and Laura Taimi. Turun yliopisto, 2014, pp. 27–35. ISBN: 978-951-29-5980-8. URL: http://urn.fi/URN:ISBN:978-951-29-5980-8.