CREAPY: A PYTHON-BASED TOOL FOR THE DETECTION OF CREAK IN CONVERSATIONAL SPEECH

{Michael Paierl, Thomas Röck}[1] , Saskia Wepner[2], Anneliese Kelterer, Barbara Schuppler

Signal Processing and Speech Communication Laboratory, Graz University of Technology
{paierl, thomas.roeck, saskia.wepner}@student.tugraz.at, anneliese.kelterer@edu.uni-graz.at, b.schuppler@tugraz.at

ABSTRACT

The annotation of creaky voice is relevant for various linguistic topics, from phonological analyses to the investigation of turn-taking, but manual annotation is a time-consuming process. In this paper, we present `creapy`, a Python-based tool to detect creaky intervals in speech signals. `creapy` does not require prior phonetic segmentation and supports the export of Praat TextGrid files, allowing for manual revision of the automatically labelled intervals. `creapy` was developed and tested using Austrian German conversational speech. It was optimised for recall to facilitate a semi-automatic annotation process, and it achieved a better performance for men's (recall: .79) than for women's voices (recall: .60).

Keywords: voice quality, creaky voice, automatic creak detection, conversational speech

1. INTRODUCTION

In this paper, we present `creapy`, a Python-based tool for automatic creak detection. Creak is a phonation type that is characterised by constriction in the glottis [1]. Identifying creaky voice is relevant for a wide range of topics in linguistics [2], such as phonological phonation type contrasts in various languages [3, 4], prosodic phrasing [5, 6], turn-taking strategies [7, 8, 9], hesitation phenomena [10], discourse structure [11], the investigation of emotion and attitude [12, 13, 14] and the sociolinguistic study of social markers [15, 16]. Automatic creak detection facilitates the identification of creak in large amounts of data and, even if manual correction may be required, it may speed up the annotation of data needed for phonetic analyses.

Apart from studying the functions of creak itself, identifying creaky voice is also relevant for an accurate extraction of fundamental frequency (F0) for other types of phonetic studies. In creaky segments, there is a significant increase of errors in the computation of F0 [17]. Either F0 cannot be detected or the detected F0 displays octave jumps due to erroneous identification of periods in the signal. Creak detection can thus support the manual correction of automatically extracted F0 by pointing to portions in the signal in which a faulty detection of F0 is likely.

Given the high relevance of creaky voice in speech science, several tools for its automatic detection have been developed. Ishi et al. [18], for instance, proposed a tool developed for conversational Japanese. Drugman et al. [19, 20] presented a creak detector that was integrated into the MATLAB-based speech processing tool COVAREP. It was developed on read and conversational speech from several languages (i.e., American English, Finnish, Swedish and Japanese). While our approach on detecting creaky voice is fairly similar to theirs, we introduce new features and do not rely fully on those proposed in [19], namely the ones used in [18, 21, 22] (H2-H1, F0, residual peak prominence, power peak parameters, inter-pulse similarity, intra-frame periodicity, energy norm, power standard deviation and zero-crossing rate).

Compared to mentioned tools, `creapy` has the advantage that it is written in the open-source programming language Python [23] and does not need manual speech segmentation. To detect creaky voice, `creapy` reads an audio signal for which it returns a binary creak decision and the underlying probability that yielded that decision for each point in time. `creapy` allows for an uncomplicated adjustment of the decision threshold as well as the parameters that are used to calculate the features. It supports the output of creak intervals in Praat [24] TextGrid files, allowing for an easy integration into a work-flow including a subsequent manual correction step. Finally, as we know that creak tends to occur more frequently in informal settings in some languages [25], we used conversational speech for its development.

2. DEVELOPMENT OF CREAPY

2.1. Materials

`creapy` was trained and tested on four conversations from the Graz corpus of read and spontaneous speech (GRASS) [26]. GRASS' conversational component contains unscripted face-to-face conver-

sations between pairs of Austrian German speakers who knew each other well (e.g., close friends, couples, family members). We found this corpus particularly suitable because we aimed at developing `creapy` on high-quality recordings of spontaneous speech, capturing a broad variety of phenomena that occur in casual conversations.

We used Praat [24] to label creaky and non-creaky intervals in 5 to 10 minutes of 4 conversations (overall 60 minutes) for each of the 8 speakers (4 women and 4 men). Intervals were labelled as being creaky by a combination of auditory analysis and visual inspection of the speech waveform (cf., [27]), spectrogram and F0 contours. For the set of non-creak labels, we chose a variety of vowels and voiced consonants, excluding voiceless fricatives, plosives and noise-like intervals, because creak only occurs in voiced segments. This resulted in 682 creak and 771 non-creak intervals to train and test the tool.

2.2. Methods

`creapy` identifies creaky voice by means of a Random Forest classifier (RFC) [28] trained on data that was labelled for the presence of creak (cf., Section 2.1). For both classification and determining feature importance, RFCs were built with 99 estimators, the default maximum depth, with a minimum samples split of 2, the square root as maximum number of features considered for splitting a node and Gini impurity measure. Given an audio file as input, `creapy` creates a list of intervals classified as creak.

2.2.1. Acoustic Features

We extracted a set of 89 acoustic features: EGEMAPSV02 features from openSMILE [29] and Cepstral Peak Prominence (CPP), which showed to be a good correlate of voice quality in many studies [30]. The features were sorted by relevance with the built-in feature importance of the RFC. The five highest ranked features were the amplitude difference of the first and second harmonic (H1H2), the Harmonics-to-noise-ratio (HNR), the fluctuation of the periodicity in time (Jitter) and amplitude (Shimmer), and the mean of the fundamental frequency (F0-mean). These features were computed for each labelled interval of speech in the training set (cf., Section 2.1). In a minority of cases, features that rely on an accurate F0 detection could not be calculated. Due to the RFC's inability to handle missing values, those had to be imputed, i.e., replaced with an actual numeric value. We chose the median of all valid results of the feature in question. This imputation step is valid for our application because `creapy` pre-eliminates unvoiced intervals prior to the actual creak detection (cf., Section 2.2.3).

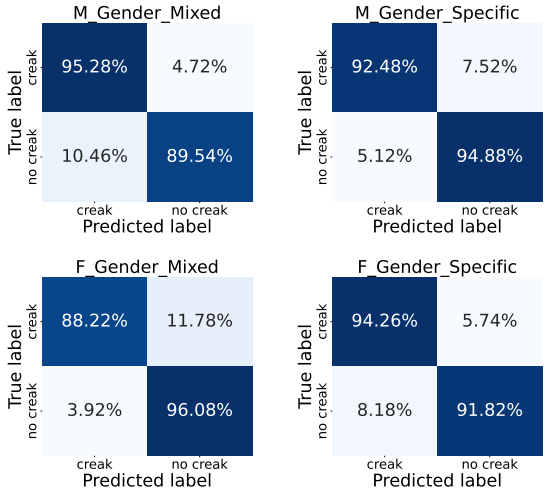

Figure 1: Confusion matrices of creak classification, averaged over speakers of the same gender with the gender-mixed models (left) and the gender-specific models (right).

2.2.2. Evaluation of the Feature Set

75% of the data was used for training and 25% for testing while performing a cross-validation procedure on all speakers. The distribution of labelled creaky and non-creaky intervals was not balanced for each speaker. Speakers were evaluated for two different classification models, one trained on speakers of mixed gender (GenMix) and one trained on speakers of the same gender (GenSpec, gender-specific). Figure 1 shows the confusion matrices of the cross-validation for each model, averaged over the speakers of each gender. Precision, F1-score and recall of approx. 90% for all different models indicates a reliable classification of creak.

2.2.3. Creak Detection

In its main application, i.e., the detection of creak, `creapy` does not need any labelled data to perform creak detection in an audio file. Figure 2 shows a schematic representation of the whole creak detection procedure. For processing the audio file, we found that using windows of 40 ms length and a frame shift of 10 ms showed good results. As creak only occurs in voiced segments, we discard voiceless or silent windows: We first calculate the short-term energy (STE) of each window, and if below 0.005, those windows are discarded as considered to contain silence. For those windows with higher STE, we calculate the zero-crossing rate (ZCR); for ZCR above 0.12, those windows are considered

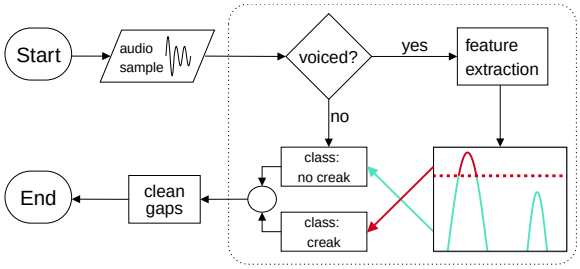

Figure 2: `creapy`'s processing steps.

voiceless and are discarded as well. This approach differs from [19], where ZCR is used as one of the features for the classifier. The five acoustic features used for creak classification (cf., Section 2.2.1) are then only extracted for the remaining intervals. If needed, missing values are imputed using the replacement value taken from the training data (cf., Section 2.2.1). Based on these acoustic features, the probability of the current interval being creaky is determined. If the probability is ≥ 0.75, the interval is classified as creak. We acknowledge that the difference between non-creaky and creaky voice is gradual rather than binary. Nevertheless, we chose this threshold because it maximised the F1-score over several test-runs with different speakers.

To create meaningful creak intervals, a post-processing step ensures creaky intervals with a minimum length of 30 ms and maximum gaps between neighbouring creak intervals of 10 ms (cf., Figure 3). Finally, the resulting creak intervals (bottom row in Figure 3) are exported to a Praat TextGrid file.

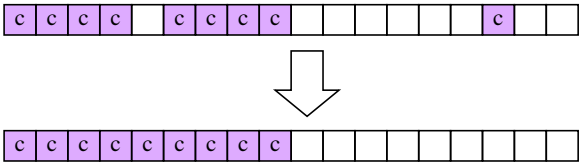

Figure 3: Schematic representation of `creapy`'s post-processing step to join neighbouring or discard very short creak intervals.

3. EVALUATION AND DISCUSSION

In this section, we present how `creapy` performs in long, continuous sound files and we compare cross-validation results from the gender-mixed (GenMix) and gender-specific (GenSpec) models. There are three possible outcomes to evaluate the detection: If the tool detected a creaky interval which overlaps with a manually labelled creak interval, this interval counts as true positive (TP). If there was no manual creak annotation at that point, the interval counts as false positive (FP). If a manually annotated creak interval was not detected by `creapy`, this counts as

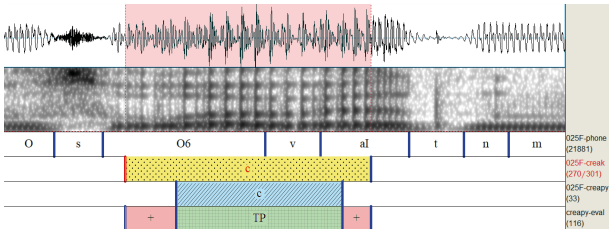

Figure 4: Example for a detected creak interval; second tier (yellow/dotted): manual label, third tier (blue/diagonal): `creapy`'s detection, fourth tier (green and red): evaluation of the detection.

false negative (FN). Figure 4 shows a TextGrid with manually annotated creak (yellow) and creak as detected by `creapy` (blue). The bottom tier shows the duration of a TP in green/checkered and the inaccuracy (i.e., time difference between manually labelled and detected intervals) in red/clean.

Table 1 shows the number of TPs, FPs, FNs, and F1- and recall-scores of the detection for the gender-specific models (GenSpec) and the gender-mixed model (GenMix). Overall, performance decreased in comparison to the evaluation of the feature set from Section 2.2.2. This decrease in performance is not unexpected, as the training material only contained manually annotated creaky and modal intervals. The continuous sound files contain a variety of speech and non-speech sounds, which were more difficult to distinguish from creak. A qualitative analysis showed that FPs often occurred in specific cases: (a) in other non-modal phonation types, e.g., pressed or breathy voice, (b) in speaker noises, such as laughter and throat clearing that may contain other kinds of non-modal phonation, (c) in modal [a], and (d) in overlapping speech when the interlocutor was audible on the speaker's audio channel. Thus, a higher performance can be expected when using `creapy` on data without overlapping speech.

Compared to the evaluation results (cf., Section 2.2.2), Table 1 shows a clearer difference in performance for speech by men vs. women. Men have a higher F1-score with relatively more FPs, and women have an overall lower F1-score and relatively more FNs. Note that this difference is *not* connected to a data imbalance, because the amount of creak labels was gender-balanced. The overall worse performance for women's speech could be related to sex-specific differences in the relationship between H1-H2 and voice quality [31]. The gender-specific model increased the number of FNs in men's data but decreased the amount of FNs in women's data. The fact that F0 octave jumps in women's creak still fall into men's modal F0 range could explain the high percentage of FNs in women's recordings processed with the gender-mixed model. With a semi-

Model	TP	FP	FN	F1	recall
M_GenMix	305	316	80	.606	.792
M_GenSpec	247	129	108	.678	.696
F_GenMix	132	55	193	.482	.406
F_GenSpec	201	103	134	.593	.600

Table 1: Cross-validation results for men and women, trained with data from all speakers (GenMix) or with gender-specific models (GenSpec).

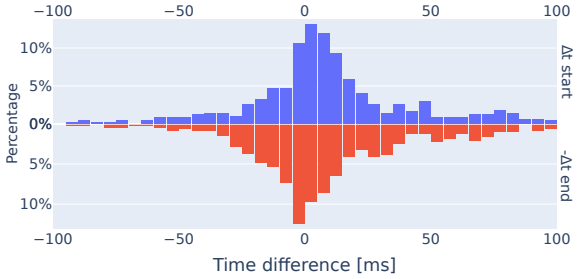

Figure 5: Absolute time difference of manual labels and detection; positive values: interval too short, negative values: interval too long.

automatic creak annotation in mind, a tool that results in fewer FNs is preferable, because it is much more time-consuming to look for creak intervals that were missed by `creapy` than to discard intervals that were incorrectly detected as creak. Therefore, we propose that the gender-specific model is better suited for creak detection for women, but not for men (cf., lower recall, despite the higher F1-score).

Our definition of TP indicates that creak was recognised correctly, but not how precisely the boundaries were detected. We thus compared the difference between start and end time of all overlapping labels (cf., Figure 4, with overlap in green/checkered and deviations in red/clean). Figure 5 shows the time differences between `creapy`'s detection and the manually annotated intervals, separately for start and end time. We inverted the sign of Δt_{end}, so that positive numbers indicate that the automatically detected interval was too short and a negative time difference indicates that the detected interval was too long. The distribution shows a shift towards positive values indicating that detected intervals tended to be too short. Some minor shifts occur inevitably due to the fixed frame shift of `creapy`. About 20 % of the time shifts are within the range of this frame shift (i.e., 10 ms). About 50 % of the intervals had an overlap of 75 % or more with manual annotations. These results indicate that while `creapy` shows a good general performance, manual adjustment of boundaries is still recommended.

`creapy` yielded results comparable to [19]'s COVAREP, a binary decision tree based on Kane et al. [22]; however, our F1-scores in Table 1 are clearly higher than their results based on [18]'s features. Yet, a detailed comparison is not possible, as their evaluation methods are not clearly identifiable.

4. HOW TO USE CREAPY

`creapy`[3] performs automatic creak detection in a continuous audio file. No previous segmentation of speech is necessary. The user can select from the following pre-trained RFC models: 1) a model trained on women only, 2) a model trained on men only, and 3) a mixed model trained on both genders (default). In the configuration file, the user can define the audio files that should be processed. While parameters from acoustic features and thresholding values (e.g., imputation strategy, ZCR and STE threshold, creak threshold) were tuned to generalise well on multiple speakers in the data we used for `creapy`'s development, these parameters can be adjusted manually in a dedicated configuration file. Detected creak-intervals are then written to a Comma-Separated-Values (`.csv`) or Praat (`.TextGrid`) file, the latter containing a new tier with the detected creak intervals. This TextGrid allows for a manual revision of the automatically created labels.

5. CONCLUSION

In this paper, we presented `creapy`, a Python-based toolkit for the detection of creak in speech signals. The best performance with a Random Forest classifier was obtained with only five acoustic features (H1H2, HNR, Jitter, Shimmer, F0-mean). In general, `creapy` achieved a good performance and we observed that a gender-mixed model performed better for men (recall: .79), and a gender-specific model performed better for women (recall: .60). Detected creak intervals had a high temporal match with manual labels. About 20 % of detected intervals overlapped almost entirely with manual labels and half of the detected intervals overlapped by 75 % or more. In the future, `creapy` could be improved with more training data of other non-modal phonation types which were difficult to distinguish from creak.

6. ACKNOWLEDGEMENTS

The work by A. Kelterer and S. Wepner was funded by grant P-32700-NB from FWF (Austrian Science Fund).

7. REFERENCES

[1] Gordon, M., Ladefoged, P. 2001. Phonation types: a cross-linguistic overview. *Journal of Phonetics* 29(4), 383–406.

[2] Garellek, M. 2022. Theoretical achievements of phonetics in the 21st century: Phonetics of voice quality. *Journal of Phonetics* 94, 101155.

[3] Garellek, M., Keating, P. 2011. The acoustic consequences of phonation and tone interactions in Jalapa Mazatec. *Journal of the International Phonetic Association* 41(2), 185–205.

[4] Kelterer, A., Schuppler, B. 2020. Phonation type contrasts and tone in Chichimec. *The Journal of the Acoustical Society of America* 147(4), 3043–3059.

[5] Belotel-Grenié, A., Grenié, M. 2004. The Creaky Voice Phonation and the Organisation of Chinese Discourse. *International Symposium on Tonal Aspects of Language: With Emphasis on Tone Languages*, 5–8.

[6] Carlson, R., Hirschberg, J., Swerts, M. 2005. Cues to upcoming Swedish prosodic boundaries: Subjective judgement studies and acoustic correlates. *Speech Communication* 46(3/4), 326–333.

[7] Ogden, R. 2001. Turn transition, creak and glottal stop in Finnish talk-in-interaction. *Journal of the International Phonetic Association* 31(1), 139–152.

[8] Lennes, M., Anttila, H. 2002. Prosodic features associated with the distribution of turns in Finnish informal dialogues. *The Phonetics Symposium*, 149–158.

[9] Włodarczak, M., Heldner, M. 2022. Contribution of voice quality to prediction of turn-taking events. *Speech Prosody 2022, Lisbon, Portugal*, 485–489.

[10] Carlson, R., Gustafson, K., Strangert, E. 2006. Cues for hesitation in speech synthesis. *INTERSPEECH*. Citeseer.

[11] Lee, S. 2015. Creaky voice as a phonational device marking parenthetical segments in talk. *Journal of Sociolinguistics* 19(3), 275–302.

[12] Gobl, C., Ní Chasaide, A. 2003. The role of voice quality in communicating emotion, mood and attitude. *Speech Communication* 40(1-2), 189–212.

[13] Horne, M. 2009. Creaky fillers and speaker attitude: Data from Swedish. In: Barth-Weingarten, D., Dehé, N., Wichmann, A. (eds), *Where Prosody Meets Pragmatics*. Emerald Group Publishing Limited, 277–288.

[14] Grichkovtsova, I., Morel, M., Lacheret, A. 2012. The role of voice quality and prosodic contour in affective speech perception. *Speech Communication* 54(3), 414–429.

[15] Yuasa, I. P. 2010. Creaky voice: A new feminine voice quality for young urban-oriented upwardly mobile American women? *American Speech* 85(3), 315–337.

[16] Mendoza-Denton, N. 2011. The Semiotic Hitchhiker's Guide to Creaky Voice: Circulation and Gendered Hardcore in a Chicana/o Gang Persona. *Journal of Linguistic Anthropology* 21(2), 261–280.

[17] Hess, W. 2012. *Pitch determination of speech signals: algorithms and devices* volume 3. Springer Science & Business Media.

[18] Ishi, C. T., Sakakibara, K.-I., Ishiguro, H., Hagita, N. 2008. A method for automatic detection of vocal fry. *IEEE Transactions on Audio, Speech, and Language Processing* 16(1), 47–56.

[19] Drugman, T., Kane, J., Gobl, C. 2014. Data-driven detection and analysis of the patterns of creaky voice. *Computer Speech & Language* 28(5), 1233–1253.

[20] Degottex, G., Kane, J., Drugman, T., Raitio, T., Scherer, S. 2014. Covarep – a collaborative voice analysis repository for speech technologies. *2014 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 960–964.

[21] Drugman, T., Kane, J., Gobl, C. 2012. Resonator-based creaky voice detection. *Thirteenth Annual Conference of the International Speech Communication Association*, 1592–1595.

[22] Kane, J., Drugman, T., Gobl, C. 2013. Improved automatic detection of creak. *Computer Speech & Language* 27(4), 1028–1047.

[23] Van Rossum, G., Drake, F. L. 2009. *Python 3 reference manual*. CreateSpace.

[24] Boersma, P., Weenink, D. 2021. Praat: doing phonetics by computer. Version 6.1.38, retrieved 2 January 2021.

[25] Kane, J., Pápay, K., Hunyadi, L., Gobl, C. 2011. On the use of creak in Hungarian spontaneous speech. *ICPhS*. Citeseer, 1014–1017.

[26] Schuppler, B., Hagmüller, M., Zahrer, A. 2017. A corpus of read and conversational Austrian German. *Speech Communication* 94C, 62–74.

[27] Keating, P. A., Garellek, M., Kreiman, J. 2015. Acoustic properties of different kinds of creaky voice. *Proceedings of ICPhS*, 2–7.

[28] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12, 2825–2830.

[29] Eyben, F., Wöllmer, M., Schuller, B. 2010. Opensmile: the Munich versatile and fast open-source audio feature extractor. *Proceedings of the 18th ACM international conference on Multimedia*, 1459–1462.

[30] Garellek, M. 2022. Theoretical achievements of phonetics in the 21st century: Phonetics of voice quality. *Journal of Phonetics* 94, 1–22.

[31] Simpson, A. P. 2012. The first and second harmonics should not be used to measure breathiness in male and female voices. *Journal of Phonetics* 40, 477–490.

[1] equal contribution

[2] corresponding author

[3] https://gitlab.tugraz.at/speech/creapy.git