INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Recognition of Creaky Voice from Emergency Calls

Lauri Tavi[1], *Tanel Alumäe*[2], *Stefan Werner*[1]

[1]University of Eastern Finland
[2]Tallinn University of Technology
lauri.tavi@uef.fi

Abstract

Although creaky voice, or vocal fry, is widely studied phonation mode, open questions still exist in creak's acoustic characterization and automatic recognition. Many questions are open since creak varies significantly depending on conversational context. In this study, we introduce an exploratory creak recognizer based on convolutional neural network (CNN), which is generated specifically for emergency calls. The study focuses on recognition of creaky voice from authentic emergency calls because creak detection could potentially provide information about the caller's emotional state or attempt of voice disguise. We generated the CNN recognition system using emergency call recordings and other out-of-domain speech recordings and compared the results with an already existing and widely used creaky voice detection system: using poor quality emergency call recordings as test data, this system achieved F1 of 0.41 whereas our CNN system accomplished an F1 of 0.64. The results show that the CNN system can perform moderately well using a limited amount of training data on challenging testing data and has the potential to achieve higher F scores when more emergency calls are used for model training.

Index Terms: creaky voice, emergency calls, convolutional networks

1. Introduction

Creaky voice, which is also known as vocal fry or laryngealisation, is to some extent a speaker-dependent phonation mode. Speakers use creaky voice for a variety of communicational reasons; previous literature has associated creak with social status [1] and voices that express boredom [2, 3], sadness or relaxedness [4]. In addition, creak has been shown to mark the ending of an utterance, for instance, in English, Estonian, and Swedish [5, 6] and turn-taking in Finnish [7].

Although creaky voice is most likely a universal speech feature and a normal mode of laryngeal production [8], some speakers tend to use it more frequently than others. Since acoustic properties and aural observation of creaky voice differ notably from those of modal voice [2], creak is a relevant voice quality considering phonetic speech science and various speech technology applications, such as natural sounding speech synthesis or speech recognition systems. Additionally, since speakers have individual ways of producing voice qualities depending on various physiological and social factors, recognition of creak can potentially be utilized for speaker recognition or classification purposes along with other speech related biometrics.

Creaky voice provides pulses that occur at a very low frequency and somewhat irregularly spaced in time [9]. Previous studies have described the physiological production of creak as follows: During creaky voice, the arytenoid cartilages may be pressed anteriorly and medially causing the posterior portions of the vocal folds to be held together [10]. The pressure on the arytenoid cartilages results in a small peak glottal opening, a long closed phase, and low glottal airflow [11, 12, 13].

As in other types of voice qualities, acoustic properties of creak vary instead of being constant, which has led to different definitions for types of creak [12, 14]. Due to the fact that creak varies considerably both between and within speakers, no robust automatic creaky voice detection systems currently exist; to the authors' knowledge, no studies have reported creak recognition accuracies of 90% especially not with low-quality recordings of conversational speech. Already existing creak recognition methods are, for instance, Ishi et al.'s creak detection algorithm [4] and widely used Voice Analysis Toolkit (https://github.com/jckane/Voice_Analysis_Toolkit) developed by John Kane and Thomas Drugman. In addition, Narendra and Sreenivasa Rao have developed a creaky detection model, which according to their study [15], outperforms the Voice Analysis Toolkit, which, in turn, outperforms Ishi's detection algorithm. However, only the Voice analysis toolkit, which includes also Ishi's detection algorithm, is freely accessible voice analysis software.

Voice quality recognition can be considered a similar task to speech recognition, even though voice quality recognition usually requires a smaller number of acoustic models and no language models. The acoustic models consist of acoustic features, such as Mel-frequency cepstral coefficients (MFCCs) that are often used with hidden Markov models (HMMs). Each HMM has typically 3-5 emitting states, and distributions of states have been traditionally represented by a Gaussian Mixture Model (GMM); however, nowadays deep neural networks (DNNs) typically have replaced GMMs in speech recognition. Although neural networks have been used for statistical modeling for decades, in recent years they have become increasingly utilized in the field of speech technology due to development of powerful GPUs. Using more computational power makes it possible to train DNN-based acoustic models on a very large scale, which has led to lower word error rate in comparison to GMMs even with small languages such as Estonian [16]. Yet, deep learning models are still rarely used for automatic voice quality detection in today's phonetic research.

As mentioned above, challenges exist both in creak's acoustic characterization and its automatic detection. Additionally, even manual labeling of creak is a more complicated task that one might expect; due to varying acoustic properties and the lack of a clear physiological or perceptual definition of transitions between voice qualities [4], even trained phoneticians' timings of creaky segment boundaries in the same speech signal may show some individual variation. This also affects evaluations of automatic systems since the recognition accuracy is often measured in comparison to manual labeling. Moreover, [17] reported that in their study increasing the amount of creaky speech samples in the training data did not improve creak classification, because heterogeneous creak increased the variance

http://dx.doi.org/10.21437/Interspeech.2019-1253

and essentially noise in the feature vectors. Thus, even though an extensive number of studies have analyzed creaky voice, many questions about the classification of creaky voice have remained open.

In this paper, we present an automatic creaky voice recognizer for emergency call recordings based on convolutional neural networks (CNNs). Detecting creak could potentially provide information about the reported emergency in relation to the speaker's emotional state. In addition, since creak is one of the most used types of voice disguise [18], a caller's unexplained permanent creaky voice is possibly an indicator of an attempt of voice disguise. The relation of creaky voice to these forensic scenarios, however, still requires more comprehensive investigations; hence, in comparison to subjective estimation of an emergency response centre (ERC) operator, an automatic creak detection would provide quantitative and more commensurate data for large scale analysis of emergency calls.

It should be noted that due to bandpass-filtering of the signal, varying connection quality, background noises and unknown callers, emergency call recordings are challenging data for automatic voice quality recognition. In addition, this kind of highly confidential data is neither commonly available nor labeled for scientific purposes. Therefore, in addition to limited amount of emergency call recordings, high-quality conversational speech recordings from laboratory-like settings was used as out-of-domain data in CNN-system training. After generation of the CNN-based creaky voice recognition system, the performance was compared to that of Voice analysis toolkit using the same authentic emergency calls.

The comparison of systems' accuracies is presented in Result section. The following section will introduce speech recordings and labeling technique.

2. Speech materials

2.1. Speech recordings

The research material consists of 30 authentic Finnish emergency call recordings (ECR30) from the Kuopio Emergency Response Centre in Northern Savonia, Eastern Finland. Additionally, the speech material includes 30 Finnish speakers from the Aalto University DSP Course Conversation Corpus (http://urn.fi/urn:nbn:fi:lb-2016051603). These 30 speakers (ACC30) produced approximately 20 utterances in 15-minute casual conversations, which were recorded during 2013–2015 using Labtec and Logitech USB headsets, PC 960 and H390. The recordings were saved in WAV and 44.1 kHz format. Only a part of the Aalto University DSP Course Conversation Corpus (i.e. 30 of 181 speakers) was used in this study because creak label files had to be created manually with [19].

The number of speakers in ECR30 is approximately 50; all emergency callers are different persons, but occasionally the same ERC operator answered different calls. The calls were made in 2016 using various unknown types of telephones and saved also in WAV format. Distributions of female and male speakers are equal for both corpora when the ERC operators are excluded. Since the recordings consist of telephone speech, they have a sampling frequency of 8 kHz. All the phone recordings were edited to approximately 20 utterances per speaker; yet, these utterances are shorter than the utterances in ACC30 since ECR30 consists of 30 approximately two-minute telephone calls. In the ECR30 recordings, callers report an ongoing emergency or suspicion of an emergency. In addition to the 30 callers, ERC operators from the same calls are included in

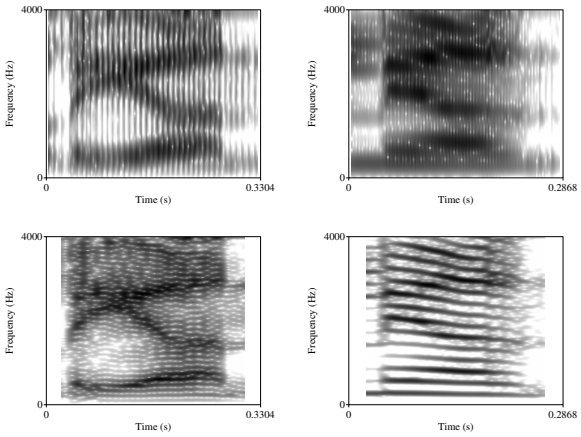

Figure 1: *Wideband (above) and narrowband (below) spectrogram of two different observations of a Finnish word /liian/, "too". The word is pronounced with creaky (left) and modal (right) voice by the same female speaker in ACC30.*

ECR30. All utterances in both corpora vary in duration from less than half a second 'yes—no' answers to almost a minute-long verbal descriptions of the situation.

2.2. Creak labeling technique

Manual labeling of creak followed a similar approach as in [5]; the decision between creaky and noncreaky voice was based on 1) auditory observation of 'crackling' voice quality, 2) irregular or low pitch contour and 3) visual examination of the spectrogram. In addition to the fact that human listeners can hear creak relatively easily, creak is also visually recognizable from spectrograms. However, it occurs differently depending on the spectrogram settings (see Figure 1).

The spectrograms in Fig. 1 show visible differences between creaky and modal phonation; for example, the wideband spectrogram reveals dispersed vertical pulses in creak, whereas in modal voice no such openings exist. This difference is shown especially in the core of the formants. On the other hand, in the narrow band spectrogram creaky voice lacks clear harmonic structure due to abnormally low f0 or even irregular periodicity. Hence, the spectrogram of creaky voice is smeary when compared to the spectrogram of modal voice where individual harmonics are clearly horizontally distinguishable.

3. Recognition of creaky voice

3.1. The amount of creak in the corpora

As mentioned in the Introduction, creaky voice is a normal mode of laryngeal production [8], but some speakers use it more frequently than others [1]. Factors such as sex and social context affect the amount of creak [6], which leads to varying prior probabilities of speakers' creak proportions. Using realistic prior probabilities is essential in statistical classification since, for instance, training a creaky recognition model with speakers, who have unusually creaky voices, will result in high number of false positives when recognizing creak from less creaky speech samples. Some evidence exists that the amount of creaky voice is correlated to the speaker's sex. For instance, increased creaky voice is more related to young female than to male speakers in American English [20], but, on the contrary, males seem to

use more creak than females in Estonian especially when they are talking to females [6]. Thus, based on the current knowledge, conclusions about the effect of speaker's sex on creakiness should treated with caution since contextual factors seem to have a different effect on the amount of creak depending on the sex.

Another interesting question raised in the literature is how emotional states affect the amount of creaky voice. An increased amount of creak has been related to boredom, relaxedness and other non-stressed conditions [2, 3, 4]. In stressful situations such as emergencies, breathiness and f0 increases and muscular tensions occur in the vocal tract [21, 22]. These effects of emotional stress are somewhat contradictory to characteristics of creak production, i.e. decreased airflow and low f0 [12]. Thus, in addition to the fact that relaxedness is related to increased creakiness, the known effects of emotional stress on voice production also suggest that emotional stress could be related to decreased creakiness.

To investigate how the amount of creak is related to a speaker's emergency conditions and sex, a two-way ANOVA was conducted with corpus (i.e. ECR30, in which emergencies are reported vs. ACC30, which contains only casual conversations) and speaker's sex as independent variables and creak proportions as dependent variables. The results of a two-way ANOVA are presented in the section Results.

3.2. CNN recognition system

We generated the automatic creaky voice recognizer based on a convolutional neural network (CNN) which consists of three alternating two-dimensional convolution and pooling layers, followed by one fully connected layer and a final softmax layer. The fully connected layer included 128 neurons. For convolutional layers, the number of filters was set to 32 for the first, 64 for the second and 128 for the third layer. The kernel sizes of the convolutional layers were 5x5, 3x3 and 3x3, respectively. The outputs of the layers were processed by max-pooling layers that down-sampled the signal by a factor of two in both dimensions. Batch normalization was performed after all pooling layers, and leaky rectified linear unit (Leaky ReLU) was used as the activation function.

CNN training and prediction was implemented in Keras [23] running on top of TensorFlow [24]. The input features were 32 mel-scale log filter-bank outputs which were extracted with Kaldi [25] from 25 ms windows using a frame shift of 10 ms. The low cutoff frequency was set to 50 Hz, while the high cutoff frequency was set to 3.5 kHz. In addition, the audio files in ACC30 were down-sampled to 8 kHz to correspond to the sampling frequency of the telephone signal.

In general, the amount of speech material in this study, i.e. approximately 4.5 hours of speech recordings, is relatively limited for training DNNs. However, currently no emergency call corpora with any voice quality annotations exist or they are inaccessible for scientific purposes. Due to the limited amount of data no computational capacity from a GPU was required for model training; hence, CNNs were trained using CPUs within 12 hours.

A total of 1645 utterances served as training data and 212 utterances served as test data. Training data consisted of ACC30 and 26 emergency call recordings. Four phone calls with nine different speakers served as test data. Additionally, an exploratory data augmentation technique was used for increasing the training data capacity; using the Praat Vocal Toolkit [26] the training data were doubled by shifting formants slightly down (0.9) and adding 50 dB of noise in the augmented ACC30 recordings to create telephone-like distortion. The aim was to produce natural sounding speech, which differs from the original speech data.

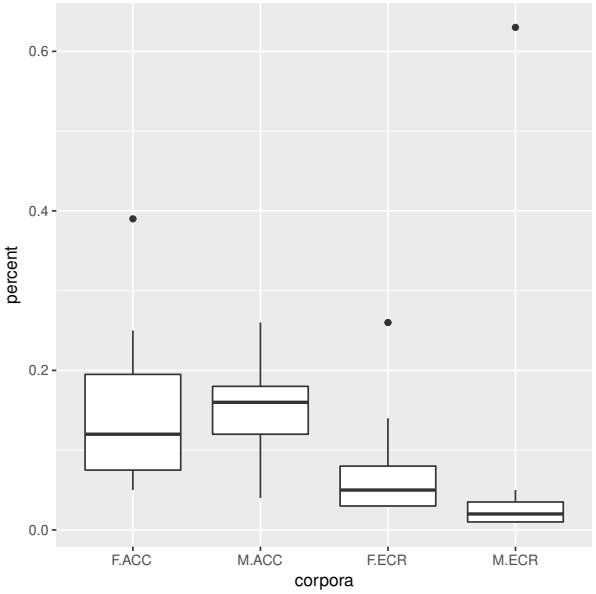

Figure 2: *Distributions of creak proportions for female (F) and male (M) speakers in ECR30 and ACC30. Creak proportions are calculated from the speech signal, excluding pauses and overlapping speech. In addition, ERC operators' utterances are excluded from the distributions. The plot was generated with R (https://www.r-project.org/).*

In the evaluation of how the automatic recognizer finds creak in an unsegmented stream of speech, this study used traditional speech recognition accuracy metrics, i.e. precision, recall and F-measure. The following (1) shows the F-measure equation.

$$2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (1)$$

These measures were defined as follows: precision is the fraction of automatically recognized creak that is actual creak. Recall, which is also known as true positive rate or sensitivity, is the fraction of correctly recognized creak from all actual creak in speech signal. Finally, a harmonic mean of precision and recall is called F-measure or F1. The recognition accuracies of the CNN-system and the Voice analysis toolkit on emergency calls will be presented in the following section.

4. Results

4.1. A two-way ANOVA

Results from the previous studies indicate that 1) male and female speakers with different native languages may have opposite tendencies for creak usage and 2) creakiness is associated with relaxed speech. The latter observation is supported by the speech data of this study: as Fig. 2 illustrates, in comparison to casual conversations in ACC30, ECR30 speakers use less creaky voice in emergency conditions. Yet, no such clear difference exists between males and females in Finnish.

To test the statistical significance of the differences in creak

proportions between the emergency-related and casual conversations and speakers' sexes, a two-way ANOVA was conducted. To exclude the outliers (see Figure 2) from the ANOVA without unbalancing the test design, one maximum value was removed from all four groups, i.e. one male and female speaker from both corpora. In addition, square root transformation was carried out on creak proportions to fulfill the normality assumption.

A two-way ANOVA supported the visual observations from Fig. 2. Analysis of variance revealed that the main effect of emergency was statistically significant; the p-value of the corpus is less than 0.001, which indicates that emergency context is associated with less creakiness in comparison to more casual conversation (see Table 1). Another main effect, sex, was not statistically significant, although the p-value was near the alpha level of 0.05 (p=0.061). Yet, the p-value for the interaction between corpus and sex was 0.004 (i.e. significant), which shows that sex is associated with creak proportions depending on the corpus variable.

Table 1: *Results from a two-way ANOVA. Statistically significant variables are shown in bold type.*

Variable	Dfs	Mean square	F-value	p-value
Corpus	1	0.391	71.755	**<0.001**
Sex	1	0.020	3.657	0.061
Corpus:Sex	1	0.049	8.980	**0.004**

The results of ANOVA support previous findings of creak's relation to tediousness and suggest that, in comparison to casual conversation, creaky voice is used significantly less when a speaker is in an emergency. On the other hand, the difference in creak proportions between sexes is not statistically significant. Although the interaction between sex and corpus passed the alpha level, the lack of clear difference between female and male speakers in two distinct data sets (i.e. ACC30 and ECR30) indicates that in the Finnish language other contextual factors of conversation, such as emotional state, affect creakiness more than speaker's sex. However, more research about creak's relation to emotional stress is required; because also other factors than sex and emotional state can affect the amount of creaky voice, so far no studies support the idea that creak levels in female or male speaker's voice could exclusively reveal information about the speaker's emotional state.

4.2. Automatic recognition

The recognition results from the CNN-system using different training procedures and the Voice analysis toolkit (VAT) are presented in Table 2. As Table 2 shows, F1, precision and recall are relatively high when only high-quality recordings from ACC30 are used for training and testing the model. Using only 20 speakers for model training and another 10 speakers for evaluation, the CNN-based system achieved F1 of 0.74 with precision of 0.76 and recall of 0.73 with conversational speech data. However, when the model was trained with ACC30 and tested with 212 utterances from four different emergency calls (ECR04), the accuracy decreased to an F1 of 0.52. The decrease of F1 was expected since emergency call recordings differ from recordings in laboratory-like settings both in their communicational (e.g. the use of creak) and technical aspects (e.g. audio quality).

When the CNN-system was trained with a combination of ACC30 and 26 emergency calls (ECR26) and tested with ECR04, F1 increased to 0.61. In addition, data augmentation (i.e. duplication of training data using formant and noise modifications) gave some improvement on the accuracy; F1, precision and recall achieved scores of 0.64, 0.60 and 0.69, respectively. In comparison to the system without data augmentation, increasing training data with speech modifications caused the system become more sensitive (i.e. higher recall), but also more vulnerable to false positives (i.e. lower precision).

Table 2: *Accuracy metrics of recognizer systems using different data sets (D=data augmentation).*

(Train)Test	F1	Precision	Recall
(ACC20)ACC10	0.74	0.76	0.73
(ACC30)ECR04	0.52	0.40	0.74
(ACC30+ECR26)ECR04	0.61	0.68	0.56
D(ACC30+ECR26)ECR04	0.64	0.60	0.69
(VAT)ECR04	0.41	0.28	0.78

All CNN-systems outperformed the VAT in creak detection from emergency call recordings: the highest recall (0.78) shows that the VAT algorithms detect actual creaky regions well, but precision of 0.28 indicates that the VAT is too sensitive for false positives with poor quality emergency call recordings. The results from aforementioned systems, which were trained using different kinds of data, show the relevance of using similar speech in model training and its actual usage.

5. Conclusion

In this study, a CNN-based creaky voice recognizer for emergency calls was generated. Considering both the limited amount of training data and the quality of testing data, the CNN-based creaky voice recognizer performed moderately well, particularly in comparison to the widely used creaky voice recognizer from the Voice analysis toolkit which achieved an F1 of 0.41 with emergency call recordings, while our CNN-system achieved 0.64 on the same test set. The results highlight the importance of using corresponding speech data in system training and its actual use. Furthermore, the results suggest that the deep learning approach is currently the most efficient method also in voice quality recognition.

Even though recognizing creaky voice is a relatively simple task for most naïve human listeners, current results from automatic recognition are still far from perfect accuracy, especially with such challenging speech data as emergency calls. An F1 of 0.64 still requires improvement but the results are promising for generating DNN-based emergency call creak recognition systems using more extensive data bases. One aim for future research is to collect more authentic emergency call recordings and create more labeled creak data using the present exploratory creak recognizer. Using more extensive labeled databases along with the proposed data augmentation technique (see 3.2), a more accurate creak detection system could be created. This would enable large-scale data analysis required to reveal the precise relation of callers' creaky voice to their different emotional states.

6. Acknowledgements

This study was supported by Jenny and Antti Wihuri Foundation (Grant No. 00160426). We would also like to thank the Emergency Response Center Administration Finland for their cooperation in providing the research material.

7. References

[1] T. Drugman, J. Kane, and C. Gobl, "Data-driven detection and analysis of the patterns of creaky voice," *Computer Speech & Language*, vol. 28, pp. 1233–1253, 2014.

[2] J. Laver, "The Phonetic Description of Voice Quality," Cambridge University Press, 1980.

[3] C. Gobl, and A. Ni-Chasaide, "The Role of Voice Quality in Communicating Emotion, Mood, and Attitude," *Speech Communication*, vol. 40, pp. 189–212, 2003.

[4] C. Ishi, K. Sakakibara, H. Ishiguro, and N. Hagita, "A method for automatic detection of vocal fry," IEEE Transactions on Audio, Speech, and Language Processing, vol. 16, no. 1, pp. 47–56, 2008.

[5] N. P. Narendra and K. Sreenivasa Rao, "Generation of creaky voice for improving the quality of HMM-based speech synthesis," *Computer speech & Language*, 42, 38–58, 2017.

[6] K. Aare, P. Lippus, and J. Simko, "Creak as a feature of lexical stress in Estonian," in Proc. INTERSPEECH, Stockholm , Sweden, 2017, pp. 1029–1033.

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

[8] H. Hollien and R. W. Wendahl, "Perceptual study of vocal fry", *The Journal of the Acoustical Society of America*, vol. 43, pp. 506–509, 1968.

[9] J. Laver, "Principle in phonetics", Cambridge University Press, 1994.

[10] M. Blomgren, Y. Chen, M. L. Ng, and H. R. Gilbert, (1998). "Acoustic, aerodynamic, physiologic, and perceptual properties of modal and vocal fry registers," *The Journal of the Acoustical Society of America*, vol. 103, pp. 2649–2658, 1998.

[11] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *The Journal of the Acoustical Society of America*, vol. 87, pp. 820–857, 1990.

[12] P. Keating, M. Garellek, and J. Kreiman, "Acoustic properties of different kinds of creaky voice," in Proceedings the 18th International Congress of Phonetic Sciences, 2015.

[13] Z. Zhang, "Vocal instabilities in a three-dimensional body-cover phonation model," *The Journal of the Acoustical Society of America*, vol. 144, pp. 1216–1230, 2018.

[14] J. Kane, T. Drugman, and C. Gobl, "Improved automatic detection of creak," *Computer Speech & Language*, vol. 27, pp. 1028–1047, 2013.

[15] N. P. Narendra and K. Sreenivasa Rao, "Automatic detection of creaky voice using epoch parameters," in Proc. INTERSPEECH, Dresden, Germany, 2015, pp. 2347–2351.

[16] A. Paats, T. Alumäe, E. Meister, and I. Fridolin, "Retrospective Analysis of Clinical Performance of an Estonian Speech Recognition System for Radiology: Effects of Different Acoustic and Language Models," *Journal of Digital Imaging*, vol. 31, pp. 1–7, 2018.

[17] A. Cullen, J. Kane, T. Drugman, and N. Harte, "Creaky voice and the classification of affect," in Proceedings of WASSS, Grenoble, France, 2013.

[18] H. J. Künzel, "Effects of voice disguise on speaking fundamental frequency," *International Journal of Speech Language and the Law*, vol. 7, pp. 150–179, 2007.

[19] P. Boersma, and D. Weenink, "Praat: doing phonetics by computer [Computer program]," Version 6.0.36, 2017, available at: http:/praat.org.

[20] N. B. Abdelli-Beruh, L. Wolk, and D. Slavin, "Prevalence of vocal fry in young adult male American English speakers," *Journal of Voice*, vol. 28, pp. 185–190, 2014.

[21] L. Tavi, "Acoustic Correlates of Female Speech Under Stress Based on /i/-Vowel Measurements," *International Journal of Speech Language and the Law*, vol. 24, pp. 227–241, 2017.

[22] C. Kirchhübel, D. M. Howard, and A. W. Stedmon. "Acoustic Correlates of Speech when Under Stress: Research, Methods and Future Directions," *International Journal of Speech Language and the Law*, vol. 18, pp. 75–98, 2011.

[23] F. Chollet, "Keras deep learning library," 2016, available at: https://github.com/fchollet/keras.

[24] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, R. Jozefowicz, Y. Jia, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, M. Schuster, R. Monga, S. Moore, D. Murray, C. Olah, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[25] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, K. Vesely, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, and G. Stemmer, "The Kaldi speech recognition toolkit," in ASRU, 2011.

[26] R. Corretge, "Praat Vocal Toolkit, " 2012, available at: http://www.praatvocaltoolkit.com/index.html