It sounds like you have a cold! Testing voice features for the Interspeech 2017 Computational Paralinguistics Cold Challenge

Mark Huckvale[1], *András Beke*[2]

[1]University College London, London, U.K.
[2]Hungarian Academy of Sciences, Budapest, Hungary
m.huckvale@ucl.ac.uk, beke.andras@nytud.mta.hu

Abstract

This paper describes an evaluation of four different voice feature sets for detecting symptoms of the common cold in speech as part of the Interspeech 2017 Computational Paralinguistics Challenge. The challenge corpus consists of 630 speakers in three partitions, of which approximately one third had a "severe" cold at the time of recording. Success on the task is measured in terms of unweighted average recall of cold/not-cold classification from short extracts of the recordings. In this paper we review previous voice features used for studying changes in health and devise four basic types of features for evaluation: voice quality features, vowel spectra features, modulation spectra features, and spectral distribution features. The evaluation shows that each feature set provides some useful information to the task, with features from the modulation spectrogram being most effective. Feature-level fusion of the feature sets shows small performance improvements on the development test set. We discuss the results in terms of the most suitable features for detecting symptoms of cold and address issues arising from the design of the challenge.

Index Terms: computational paralinguistics, cold, respiratory tract infection, voice

1. Introduction

The goal of the Interspeech 2017 Cold Challenge was to identify speakers with upper-respiratory tract infections from short speech recordings. The training and testing corpus (URTIC) was provided by the Institute of Safety Technology, University of Wuppertal, Germany and consisted of recordings of 630 subjects made in quiet rooms. Each speaker completed a health questionnaire (WURSS24) [1] that contains 22 seven-point Likert scale questions related to symptoms of the common cold. Speakers with a mean score greater than or equal to 6 were classed as having a cold, others as not having a cold. The individual recordings were then divided into 28 652 short (3s-10s) sections and partitioned into training, development and tests sets, with no speaker being present in more than one set. For further details of the corpus, please see [2].

In this paper, we build on previous approaches we have investigated for the classification of other paralinguistic properties of the voice: for Cognitive Load [3], for Fatigue [4], and for Laughter [5]. Our strategy has been to create well-motivated feature sets that summarise temporal, spectral and modulational properties of each recording that are then used to train support-vector machine (SVM) or deep-neural network (DNN) classifiers. Section 2 of the paper investigates some previous voice features used in detecting changes in speaker health, which motivates our choice of four different voice feature sets. Section 3 describes how we extracted the voice features, and Section 4 describes how we built SVM and DNN classifiers. Sections 5 and 6 present the performance of the features singly and in combination on the challenge development and test sets, and discuss the outcomes. The paper concludes with speculation about the reasons for performance variation.

2. Choosing Voice Features

2.1. Features for detecting health changes

Since speaking engages critical respiratory and neurological functions, speech is likely to be affected by almost any condition which affects the health of the individual, if only because of the added stress on the body. In recent years we have seen reported many health issues that supposedly affect speech including: acid reflux, adenoid cystic carcinoma, asthma, brain cancer, bronchitis, cleft palate, dementia, emphysema, hay fever, multiple sclerosis, multiple system atrophy, nerve damage to muscles of the vocal cords, noncancerous growths (polyps, nodules, cysts, granulomas, papillomas or ulcers) on the vocal cords, Parkinson's disease, stroke, and throat cancer. Most of these have yet to be investigated using computational paralinguistics techniques, but there are some studies which have looked at what changes such health issues have on speech which we can use to guide the choice of voice features.

From the spectral domain, features such as formant frequencies and bandwidths, and other sub-band measures have been shown to vary due to Parkinson's [6] and asthma [7].

Time-domain features based on pitch (f0 mean, f0 variation, etc.) or voice quality (jitter, shimmer, HNR, etc.) have been widely reported as changing due to brain and mental health disorders such as depression, schizophrenia, and Parkinson's [8, 9, 10, 11, 12].

Changes in spectral envelope shape and slope are widely reported as a consequence of voice pathology [13] or Parkinson's [14, 15]. The use of mel-frequency cepstral coefficients (MFCC) to describe the short-time speech spectral envelope is very common.

Combinations of time-domain and frequency domain features have also been used for rating the severity of Parkinson's [16].

For studying the effect of fatigue on voice, Baykaner *et al.* used a combination of time-domain, spectral domain and modulation domain features [4].

2.2. Features for detecting common cold

The common cold has a number of physiological effects which may have some effect on the voice. Commonly reported symptoms include: cough, hoarseness, sore throat, nasal obstruction, sneezing, nasal leakage and nasal stuffiness [17]. However studies show a great deal of variability in how the common cold affects individuals. Some of this variability is to do with differences between the viruses causing the disease, some to do with the fact that the frequency of symptoms changes as the disease progresses within an individual and some to do with differences in the resilience of individuals themselves [17].

In trying to detect presence of the common cold from the voice, there are few studies that can be called on for advice. Those that exist seem to be related to issues of whether speaker recognition systems are affected by having a cold. For example Tull and Rutledge [18] looked at the effect of common cold symptoms on the vowel formant frequencies of 10 speakers. They saw reductions in formant frequencies in the cold voice. In a separate investigation Tull *et al.* [19] looked at the effect of cold symptoms on cepstral coefficients from digit recordings. They showed that the differences between normal and cold speech were mainly in the lower coefficients c2 and c3.

A few studies parallel the cold challenge task. Barry *et al.* [20] used MFCC and LPCC features and a neural network classifier to identify the presence of coughs. For the same application, Matos *et al.* [21] used MFCC features and HMM modelling. Larson *et al.* [22] used PCA features constructed from an FFT spectrogram together with a random forest classifier to identify coughs in speech samples.

The cold challenge baseline feature set is based on the OpenSMILE features [23] which includes a wide variety of temporal and spectral measures together with functionals that summarise these parameters over each sample.

2.3. Choosing Voice Features

In this study, we chose to evaluate four sets of features which broadly relate to the types of features found in previous studies on health and cold detection from speech:

a) **Voice Quality features** based on variations in the periodic temporal structure of the speech signal, such as pitch, irregularity, breathiness and effectiveness. These features are motivated by expected changes in larynx operation with symptoms of cold.
b) **Vowel Spectral features** based on differences in the resonance characteristics of sonorant sounds. These features are motivated by expected changes in vocal tract shape, use or obstruction of the nasal cavities caused by cold symptoms.
c) **Spectral Modulation features** based on changes in the character of amplitude modulations in different frequency bands. These features are motivated by expected changes in the effectiveness of glottal excitation to cause modulations at different frequencies, changes in vocal tract damping, and changes to articulatory quality and speaking rate. These changes might occur over a wide range of modulation frequencies.
d) **Distribution of Spectral Envelope features** based on changes to the probability distribution of spectral envelopes. These features are motivated by expected changes in the relative frequency of sound elements in the recordings, these might include the loss of nasal segments or added sounds such as sniffs, coughs & groans.

3. Extraction of Voice Features

3.1. Voice Quality Features (VOI)

To characterize voice quality, we combined a number of features generated by a mixture of existing toolkits. The occurrence of creaky voice was measured using the Voice Analysis Toolkit [24]. The Peak Slope measure [25] calculated from a wavelet-based decomposition of the speech signal into octave bands was used to differentiate breathy and tense voice qualities. To these features we added the following parameters: time domain features of zero-crossing rate, energy, and entropy of energy; and spectral domain features of spectral centroid, spectral entropy, spectral flux, spectral rolloff, MFCCs, harmonicity, pitch, and a chroma vector (12-dimensional representation of the spectral energy). These short-time measurements were then summarised for each recording using functionals of mean, standard deviation, skewness, and kurtosis to derive a feature vector of 156 values.

3.2. Vowel Spectral Features (VOW)

To extract vowel spectra, the signals were first processed by a German phone recognizer (MAUS [26]) to identify the vocalic regions. 12 MFCC coefficients plus energy, deltas and delta-deltas were then extracted over each vowel segment. The distribution of the MFCC coefficients across the vocalic segments in each recording was then described by 4 functional parameters: mean, standard deviation, skewness and kurtosis to generate a 156 feature vector.

3.3. Spectral Modulation Features (MOD)

The modulation spectrogram is calculated from 18 third-octave sub-bands of the signal. The signal is passed through a filterbank of 4th order Butterworth filters between 125Hz and 6350Hz. The normalized absolute amplitude is then taken in each channel and the modulation spectrum calculated as an average of a series of FFTs applied to 500ms Hamming-windowed sections of the envelope overlapped by 250ms. The modulation amplitudes are then log compressed and modulation frequencies up to 500Hz preserved, see Figure 1.

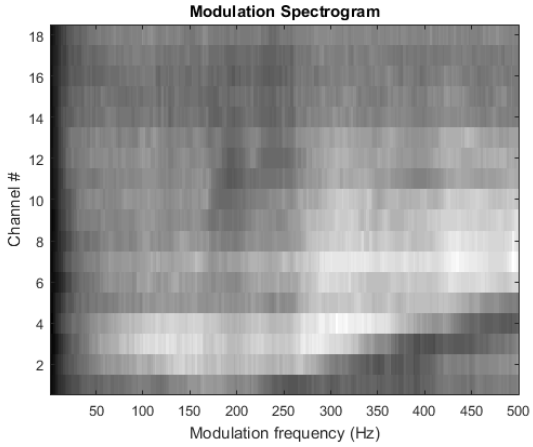
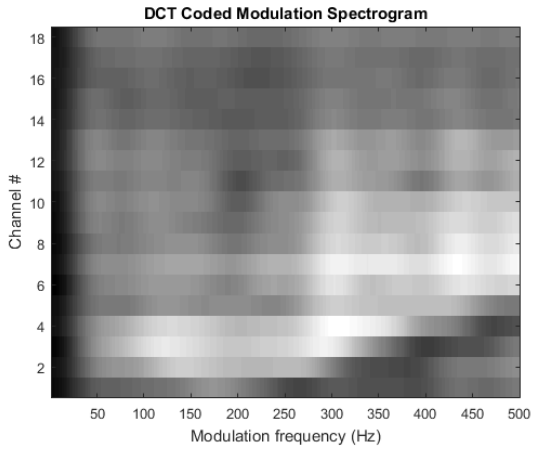


Figure 1 Modulation Spectrogram

The modulation spectrogram in the modulation frequency range of 0-500Hz is then compressed by taking the first 16 coefficients of the Discrete Cosine Transform for each channel, to generate an 18x16=288 feature vector. The effect of the data reduction can be seen in the reconstructed modulation spectrogram in Figure 2.

Figure 2 DCT Coded Modulation Spectrogram



3.4. Distribution of Spectral Envelope features (GPPS)

To capture differences in the recordings as to the relative frequency of spectral events, MFCC parameters were first extracted for each recording. These consisted of 19 coefficients plus energy and their deltas computed every 10ms. These were then mean and variance normalised per recording. A universal background model (UBM) was then constructed from the whole training set using a GMM with 512 mixtures. To extract a feature vector from each recording, the posterior probabilities of the 512 mixtures are computed from the occupancy counts of the UBM. This vector then reflects the relative frequency of 512 spectral "events" in the signals, in an analogous fashion perhaps to the "bag of audio words" used in the challenge baseline system [2].

4. Classifier construction

4.1. Normalisation

Our previous computational paralinguistics studies have shown the importance of feature normalisation to aid learning [3, 4]. Without normalisation, learning may be affected by large differences in dynamic range across features, poor distributional properties or outlier values. In this study we compared three types of normalisation:

a) Uniform normalisation: where all features are linearly scaled to the range 0-1. A possible problem with this approach is that a few outliers can significantly affect the mapping.
b) Z-score normalisation: where every feature value is measured as a distance from the mean of all feature values in units of standard deviation. This reduces the effects of outliers on the normalisation mapping.
c) Gaussianisation: where the rank of each feature value among a sorted list of available feature values is used to extract quantiles from the cumulative normal distribution [27]. This enforces a Gaussian distribution on every feature, while losing information about their absolute values.

4.2. Sample Balancing

The cold challenge corpus is unusual in a number of respects: (i) only a minority of speakers (one third) had symptoms of a severe cold at the time of recording, (ii) more extracts were chosen from non-cold subjects, so that the number of audio samples of speakers with cold was only ~18% in the training and development sets; and (iii) speakers were not repeated across training, development and test sets. We foresaw that these issues may cause problems for training an effective cold classifier, since the classifier may become too dependent on the majority (non-cold) class, or may learn properties of cold that were too specific to the speakers in the training set (at worst, the classifier may just become a speaker recognizer for those speakers in the training set that have a cold).

To address these problems we investigated the synthetic generation of new training samples from the minority class. The idea was (i) to balance the number of non-cold and cold samples, and (ii) to create "new" cold speakers from mixtures of old speakers. In this study we looked at three approaches:

a) Unbalanced: no change to the training samples.
b) SMOTE balanced: synthesis of new samples of the cold class using the SMOTE procedure [28] in which each new sample is computed as an admixture of two randomly chosen samples from the minority class.
c) ADASYN balanced: synthesis of new samples of the cold class using the ADASYN procedure [29]. In this procedure new minority samples are only generated in the area of the vector space where the density of minority class vectors is low. To measure the neighbourhood density of any vector a k-nearest neighbour measure is used based on Euclidean distances.

4.3. SVM Classifier

The LIBLINEAR package [30] was used to train and test SVM classifiers. Feature vector normalisation of uniform, z-score and gaussianisation was explored. Balancing was chosen from Unbalanced, SMOTE balanced and ADASYN balanced. Variations in the "complexity" parameter C were explored in powers-of-ten steps from 1e-6 to 10. A linear regression fit was used to obtain posterior probabilities.

4.4. Neural network classifier

The Microsoft Cognitive Toolkit CNTK [31] was used to train and test deep neural network classifiers. Feature vector normalisation of uniform, z-score and gaussianisation was explored. Balancing was chosen from Unbalanced, SMOTE balanced and ADASYN balanced. Network nodes types of Sigmoid, Tanh and Rectified Linear were explored. Networks had 1, 2 or 3 hidden layers, of 25, 50 or 100 nodes in each layer. The output layer had two SoftMax nodes to represent class probabilities. A learning rate of 0.5 stepping down to a rate of 0.1 over 30 learning epochs was used in all tests, with a stochastic gradient descent learning algorithm on a cross entropy measure applied in mini-batches of 100.

4.5. Performance measurement

The challenge performance measure is unweighted average recall (UAR), that is the labelling accuracy assuming not-cold and cold detection as equally important. To calculate UAR for our systems, we applied the classifier trained on the training

set to the development set obtaining a list of posterior probabilities for the cold class for each sample. Using the correct labelling a threshold value was then chosen to find a probability value such that the proportion of cold samples labelled as non-cold was approximately similar to the proportion of non-cold samples labelled as cold. This is the equal-error rate threshold.

5. Results

5.1. Development Set

Since our main objective in this study was to explore the difference between feature sets for the detection of symptoms of cold in the voice, we report below the performance of the best system configurations only.

Table 1 Best development set performance (UAR%)

Feature set	Best SVM	Best DNN
OpenSMILE Baseline	64.00	-
VOI	66.34	65.58
VOW	66.47	65.48
MOD	**67.95**	**67.95**
GPPS	66.07	65.58
VOI+VOWEL	68.37	66.59
VOI+MOD	64.88	68.13
VOI+GPPS	67.34	67.32
VOW+MOD	67.05	70.02
VOW+GPPS	**69.03**	69.11
MOD+GPPS	67.36	**70.97**

The best single feature set using the SVM classifier was MOD with z-score normalisation, ADAS balancing and C=0.1. The best single feature set using the DNN classifier was also MOD with z-score normalisation, using sigmoid nodes in 100:100:100 layers, but no balancing.

The best feature fusion using the SVM was VOW+GPPS with z-score normalisation, no balancing and C=1e-6. The best feature fusion using the DNN was MOD+GPPS with z-score normalisation, rectified linear nodes in 100:100:100 layers and no balancing.

5.2. Test Set

To build a system for evaluation on the test set, the system that performed best on the development set was retrained using the whole training and development sets as training data, and with the same set of configuration and learning parameters. Test set performance of the system that performed best on the development set is shown in Table 2, together with the best performing single classifier reported in the baseline

Table 2 Test Set Performance

System	UAR %
Baseline: **SVM + OpenSMILE features**	70.2
Best performing development set system: **DNN + MOD + GPPS features**	62.1

Although our best system considerably outperformed the baseline on the development set, performance is considerably worse on the test set. This may be due to some over-fitting of the best system to the development set, extreme sensitivity of the system to the choice of configuration parameters or some other discrepancy between audio samples in the corpus partitions.

6. Discussion

The outcomes of our evaluation are as follows. In terms of normalisation strategy, only small effects were seen, but overall z-score normalisation did provide the best UAR scores on this task.

In terms of balancing, we found little evidence that SMOTE or ADASYN balancing improved the UAR scores. In some configurations we saw that balancing even had adverse effects on the ability of the classifiers to learn the task.

In terms of choice of classifier, the two classifiers obtained rather similar performance for the different voice features.

In terms of classifier configuration, good performance was obtained on occasions with a wide range of SVM complexity parameters, although performance could vary by as much as ±5% for one training set across complexity settings. The same behaviour could be found for network configurations, with both Sigmoid and Rectified Linear units giving good performance on different training sets, but with a considerable range of performance figures over different network configurations. Overall the implication is that this task is very sensitive to system configuration, which may explain poor test set performance.

In terms of the best features sets, while all four proposed feature sets performed better than the baseline OpenSMILE feature set on the development set, the modulation spectrogram features were the best single set. This may be because this set captures voice changes at both high and low modulation frequencies, relevant to the symptoms of cold affecting excitation, resonance and rhythm of the speech.

7. Conclusions

The 2017 Computational Paralinguistics Cold challenge was particularly difficult for a number of reasons. The fact that the training set was based on a large number (10 000) of extracts from a few (210) speakers of which only a minority had a cold made the training of a classifier sensitive to cold rather than to speaker difficult. Since we have already noted that the symptoms of cold vary considerably across virus, individual and time [17], it is likely that even the speakers in the training set with a cold do not form a homogeneous group. The task would have been much easier if longer segments of speech were available, if duplicate recordings of the same speaker were marked, or if recordings of the same speakers with and without cold were available. In the last case, it would have allowed the training of a joint factor model to separate out the effects of cold from the effects of speaker identity, much as the same approach is used in speaker recognition to separate out effects of channel [32].

8. Acknowledgements

The work conducted here was supported in part by a European Space Agency grant: Embedded Psychological Support Integrated for Long duration missions – EPSILON, (phase 1 - VULCAN). Thanks to the organisers of the Interspeech 2017 Computational Paralinguistics Challenge for making this study possible.

References

[1] Barrett, B., Brown, R. L., Mundt, M. P., Thomas, G. R., Barlow, S. K., Highstrom, A. D., and Bahrainian, M. "Validation of a short form Wisconsin upper respiratory symptom survey (WURSS-21)". *Health and Quality of Life Outcomes*, 7(1), pp. 76, 2009.

[2] Schuller, B., Steidl, S., Batliner, A., Bergelson, E., Krajewski, J., Janott, C., Amatuni, A., Casillas, M., Seidl, A., Soderstrom, M., Warlaumont, A., Hidalgo, G., Schnieder, S., Heiser, C., Hohenhorst, W., Herzog, M., Schmitt, M., Qian, K., Zhang, Y., Trigeorgis, G., Tzirakis, P., Zafeiriou, S., "The INTERSPEECH 2017 Computational Paralinguistics Challenge: Addressee, Cold & Snoring", in *INTERSPEECH 2017 – 18th Annual Conference of the International Speech Communication Association*, August pp. 20–24, Stockholm, Sweden, Proceedings, 2017.

[3] Huckvale, M., "Prediction of Cognitive Load from Speech with the VOQAL Voice Quality Toolbox for the InterSpeech 2014 Computational Paralinguistics Challenge", *Proc. Interspeech* 2014, Singapore, 2014.

[4] Baykaner, K. R., Huckvale, M., Whiteley, I., Andreeva, S., and Ryumin, O., "Predicting Fatigue and Psychophysiological Test Performance from Speech for Safety-Critical Environments". *Frontiers in Bioengineering and Biotechnology*, 3, 2015.

[5] Gosztolya, G., Beke, A., Neuberger, T., Tóth, L. "Laughter Classification Using Deep Rectifier Neural Networks with a Minimal Feature Subset", *Archives of Acoustics* Vol. 41, No. 4, pp. 669–682, 2016.

[6] Bang, Y., Min, K., Sohn, Y., Cho, S., "Acoustic characteristics of vowel sounds in patients with Parkinson disease." *NeuroRehabilitation* 32.3, pp. 649–654, 2013.

[7] Sonu, R., "Disease detection using analysis of voice parameters." Int. *J. Comput. Sci. Commun. Technol*, 4(2), 2012.

[8] Michaelis, D., Fröhlich, M., Strube, H., "Selection and combination of acoustic features for the description of pathologic voices." *Journal of the Acoustical Society of America* 103, pp. 1628, 1998.

[9] Cohen, A., Alpert, M., Nienow, T., Dinzeo, T., Docherty, N., "Computerized measurement of negative symptoms in schizophrenia." *Journal of psychiatric research* 42, pp. 827–836, 2008.

[10] Alpert, M., Pouget, E., Silva, R., "Reflections of depression in acoustic measures of the patient's speech." *Journal of affective disorders* 66, pp. 59–69, 2001.

[11] Proença, J., Veiga, A., Candeias, S., and Perdigão, F., "Acoustic, Phonetic and Prosodic Features of Parkinson's disease Speech." In *STIL-IX Brazilian Symposium in Information and Human Language Technology,* 2nd Brazilian Conference on Intelligent Systems (BRACIS 2013), Fortaleza/Ceará, Brazil, 2013.

[12] Kliper, R., Portuguese, S., and Weinshall, D., "Prosodic Analysis of Speech and the Underlying Mental State". In *International Symposium on Pervasive Computing Paradigms for Mental Health* (pp. 52–62). Springer International Publishing, 2015.

[13] Arias-Londono, J. D., Godino-Llorente, J. I., Sáenz-Lechón, N., Osma-Ruiz, V., and Castellanos-Domínguez, G. "Automatic detection of pathological voices using complexity measures, noise parameters, and mel-cepstral coefficients". *IEEE Transactions on Biomedical Engineering*, 58(2), pp. 370–379, 2011.

[14] Benba, A., Jilbab, A., and Hammouch, A., "Detecting Patients with Parkinson's disease using Mel Frequency Cepstral Coefficients and Support Vector Machines." *International Journal on Electrical Engineering and Informatics*, 7(2), pp. 297, 2015.

[15] Shirvan, R. A., and Tahami, E., "Voice analysis for detecting Parkinson's disease using genetic algorithm and KNN classification method." In *Biomedical Engineering* (ICBME), 2011 18th Iranian Conference of IEEE, pp. 278–283, 2011.

[16] Tsanas, A., Little, M. A., McSharry, P. E., Spielman, J., and Ramig, L. O., "Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease." IEEE Transactions on Biomedical Engineering, 59(5), pp. 1264–1271, 2012.

[17] Tyrrell, D. A. J., Cohen, S., and Schilarb, J. E., "Signs and symptoms in common colds." *Epidemiology and infection*, 111(01), pp. 143–156, 1993.

[18] Tull, R. G., and Rutledge, J.C., "'Cold Speech' for Automatic Speaker Recognition." Acoustical Society of America 131st Meeting Lay Language Papers, 1996.

[19] Tull, R. G., Rutledge, J.C., and Larson, C.R., "Cepstral analysis of 'cold-speech' for speaker recognition: A second look". *Diss*. ASA, 1996.

[20] Barry, S. J., Dane, A. D., Morice, A. H., and Walmsley, A. D. "The automatic recognition and counting of cough." *Cough*, 2(1), 8, 2006.

[21] Matos, S., Birring, S. S., Pavord, I. D., and Evans, H., "Detection of cough signals in continuous audio recordings using hidden Markov models." *IEEE Transactions on Biomedical Engineering*, 53(6), pp. 1078–1083, 2006.

[22] Larson, E. C., Lee, T., Liu, S., Rosenfeld, M., and Patel, S. N., "Accurate and privacy preserving cough sensing using a low-cost microphone." In *Proceedings of the 13th international conference on Ubiquitous computing* (pp. 375–384). ACM, 2011.

[23] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent developments in openSMILE, the Munich open-source multimedia feature extractor," in *Proceedings of ACM MM 2013 Barcelona, Spain*, ACM, pp. 835–838, 2013

[24] Kane, J., and Gobl, C., "Evaluation of glottal closure instant detection in a range of voice qualities." *Speech Communication*, 55(2), pp. 295–314, 2013.

[25] Kane, J., and Gobl, C., "Identifying Regions of Non-Modal Phonation Using Features of the Wavelet Transform." In *Interspeech 2011*, pp. 177–180, 2011.

[26] Kisler, T., Schiel, F., and Sloetjes, H., "Signal processing via web services: the use case WebMAUS." *In Digital Humanities Conference*, 2012.

[27] Chen, S., Gopinath, R., "Gaussianization" *Proc. NIPS 2000,* Denver Colorado, 2000.

[28] Chawla, N., Bowyer, K., Hall, L. and Kegelmeyer, W., "SMOTE: Synthetic Minority Over-sampling Technique." *Journal of Artificial Intelligence Research*. 16, pp. 321–357, 2002.

[29] Wacharasak S., "smotefamily: A Collection of Oversampling Techniques for the Class Imbalance Problem Based on SMOTE:" https://CRAN.R-project.org/package=smotefamily 2016

[30] Fan, R, Chang, K., Hsieh, C., Wang, X. and Lin, C., "LIBLINEAR: A library for large linear classification", *Journal of Machine Learning Research* 9, pp1871-1874, 2008

[31] CNTK toolkit: https://www.microsoft.com/en-us/research/product/cognitive-toolkit/

[32] Kenny, P., Ouellet, P., Dehak, N., Gupta, V., & Dumouchel, P., "A study of interspeaker variability in speaker verification" *IEEE Transactions on Audio, Speech, and Language Processing*, 16(5), 980-988, 2008