INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

A robust Voiced/Unvoiced phoneme classification from whispered speech using the 'color' of whispered phonemes and Deep Neural Network

G. Nisha Meenakshi, Prasanta Kumar Ghosh

Electrical Engineering, Indian Institute of Science (IISc), Bangalore, Karnataka- 5600012, India
gnisha@ee.iisc.ernet.in, prasantg@ee.iisc.ernet.in

Abstract

In this work, we propose a robust method to perform frame-level classification of voiced (V) and unvoiced (UV) phonemes from whispered speech, a challenging task due to its voiceless and noise-like nature. We hypothesize that a whispered speech spectrum can be represented as a linear combination of a set of colored noise spectra. A five-dimensional (5D) feature is computed by employing non-negative matrix factorization with a fixed basis dictionary, constructed using spectra of five colored noises. Deep Neural Network (DNN) is used as the classifier. We consider two baseline features-1) Mel Frequency Cepstral Coefficients (MFCC), 2) features computed from a data driven dictionary. Experiments reveal that the features from the colored noise dictionary perform better (on average) than that using the data driven dictionary, with a relative improvement in the average V/UV accuracy of 10.30%, within, and 10.41%, across, data from seven subjects. We also find that the MFCCs and 5D features carry complementary information regarding the nature of voicing decisions in whispered speech. Hence, across all subjects, we obtain a balanced frame-level V/UV classification performance, when MFCC and 5D features are combined, compared to a skewed performance when they are considered separately.

Index Terms: Voiced and Unvoiced whispered phonemes, Classification, Non-negative matrix factorization

1. Introduction

Whispered speech is typically produced in private as well as in pathological conditions, such as laryngectomy [1]. Whispered speech lacks pitch due to the absence of vocal folds vibrations during its production [2] and hence, is voiceless and less intelligible compared to neutral speech. Therefore, several attempts have been made in order to reconstruct neutral speech from whispered speech [1, 3, 4, 5]. To ensure a natural sounding reconstructed speech, it is essential to perform, both, the estimation and appropriate incorporation of pitch. This process typically requires a voiced (V) and unvoiced (UV) decision from the whispered speech. Hence, automatic classification of V and UV phonemes from whispered speech becomes vital. Since, V and UV phonemes are characterized by the presence and absence of pitch, respectively, the task of classifying V/UV phonemes[1] from a speech that is typically voiceless, is a challenging one, although they are perceptually discriminative [6].

There exist several algorithms to reconstruct neutral speech from whispers, by a direct estimation of the pitch contour without an intermediate V/UV classification step. These include prediction of pitch from spectral features via a statistical model [5, 7] and estimation of pitch as a function of formants [8]. These procedures typically suffer from an unnatural prosodic contour and formant estimation errors, respectively. Therefore, to obtain a natural pitch contour, there is a need to first predict if a given frame of whispered speech is V or UV.

The task of classifying V/UV phonemes from whispered speech has been addressed in the past. Sharifzadeh *et al.* used a frame energy based technique to classify V and UV frames from pathological whispered speech [1]. This work requires patient specific manually chosen thresholds for the classification task. An approximate measure of voicing, the energy ratio between higher and lower frequencies, was employed by Morris *et al.* [3], while a formant count procedure was adapted by Ahmadi *et al.* [4]. The gender specific shift of formant frequencies [9] and the noise-like nature of whispered speech [8], could lead to a poor performance of such approximations and formant estimation.

In our work, we exploit the noise-like nature of whispered speech to obtain voicing cues, by attempting to determine the 'color' of whispered V and UV phonemes. In order to do so, we hypothesize that a whispered speech spectrum can be represented as a linear combination of spectra from colored noises. In the proposed method, we consider five colored noises, namely, Violet, Blue, White, Pink and Brown. We consider frame-level V/UV classification. Interestingly, a Deep Neural Network (DNN) classifier trained on the features from the coefficients corresponding to a dictionary of colored noises, computed using Non-negative Matrix Factorization (NMF), outperforms a DNN classifier trained on those from a dictionary learnt directly from the data. A combination of spectral features and those computed using the color noise dictionaries, is found to yield an equally good performance on both the V and UV phoneme classes, across and within seven subjects, in comparison to the skewed performance exhibited by the two baseline schemes considered in the study. We begin with an interpretation of the 'color' of whispered V and UV phonemes in Section 2.

2. The 'color' of whispered V/UV phonemes

It is known that the V speech segments have a steeper spectral slope compared to that of the UV segments, in neutral speech [10, 11]. The noise-like whispered speech may not always follow a trend in the steepness of the spectral slopes between V and UV phonemes, similar to that of neutral speech. Therefore, we analyze if the noise-like whispered speech spectrum could be represented as a linear combination of colored noises. We hypothesize that the combination of the spectra of the colored noises, in order to represent a V spectrum would be different from that to represent a UV spectrum. As a preliminary attempt to validate this hypothesis, in this work, we consider the spectrally flat white Gaussian noise, along with two noises with decreasing and two noises with increasing spectral slopes. Blue

[1] Although 'voiced' phonemes are not voiced (lack pitch) while whispering, we still address them as 'voiced' for convenience.

http://dx.doi.org/10.21437/Interspeech.2017-1388

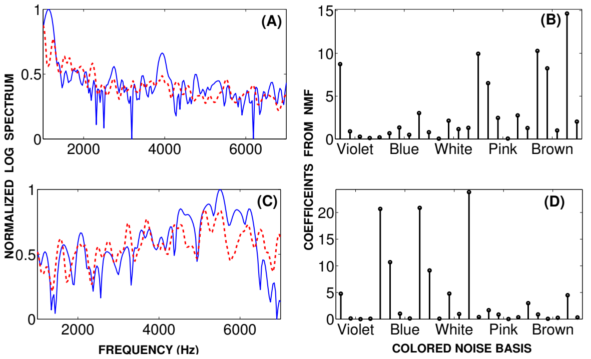

Figure 1: *Original spectra, in blue continuous line, and the reconstructed spectra, in red dashed line, (using five CNBVs for each of the five colored noises in NMF) for one V, (A), and one UV, (C), spectra with corresponding coefficients, (B) and (D), respectively.*

and Violet noises are characterized by an increasing spectral slope of f and f^2, respectively, while Pink and Brown noises are characterized by decreasing slopes of $1/f$ and $1/f^2$, respectively, where f denotes frequency. We then construct a colored noise dictionary (CND) using spectra from each of the five noises and employ the NMF technique [12] to estimate the contribution of each of these colored noise basis vectors (CNBV).

We begin by considering a whispered speech utterance of length N samples, denoted by $x[n]$, n=0 . . . $N-1$. We then obtain the spectrogram, $P \geq 0$, using a window of length N_w and a shift of N_{sh} samples. Therefore, we have a non-negative matrix P of dimensions $N_F \times N_t$ and a fixed dictionary W_N of dimension $N_F \times r$. Using NMF, we compute a matrix of coefficients H^*, of dimensions $r \times N_t$, by solving the following Euclidean norm ($\|.\|_2$) minimization problem,

$$H^* = \underset{H\geq 0}{\arg\min} \|P - W_N \times H\|_2, \qquad (1)$$

where, $W_N \geq 0$ is the fixed CND with rank r and H^* is the corresponding coefficient matrix.

Fig. 1 shows the original spectra (from 1 to 7kHz) and the reconstructed spectra (a column of $W_N \times H^*$ corresponding to the concerned V/UV spectrum) for one V frame (A) and one UV frame (C). The contribution of the CNBVs (a column of H^* corresponding to the concerned V/UV spectrum) is shown in (B) and (D), respectively. It can be seen that Brown and Pink noises with a decreasing spectral slope contribute more to the V spectra while noises such a White, Violet and Blue with zero or an increasing slope contribute more to the UV spectra. We make two important observations: 1) The V and UV phonemes are not completely characterized by one noise of a particular spectral slope, but are characterized by a combination of several noises, 2) The CNBVs contributing to V are different from those to UV.

As seen from Fig. 1, the reconstructed spectrum does not exactly match with the original spectrum. Therefore, we analyze if these fixed sets of colored noises indeed capture the trend in the spectrum of V and UV phonemes. For this, we compute the frame-wise spectral slope [13] of the original and the reconstructed spectra, using the whispered speech data from four female (F1-F4) and three male (M1-M3) subjects ($\sim$ 35000 frames per subject) and compute the histogram of their difference, as shown in Fig. 2. We observe only a small difference of the order 10^{-3}, indicating that the CNBVs from these five noises could be sufficient to capture the slope of the whispered V and UV phonemes. It is known that the whispered speech is characterized by a lower spectral tilt [11, 13], compared to the neutral speech. The lower values of spectral slope may not effectively capture the differences between whispered V/UV phonemes. Therefore, we hypothesize that the pattern of the relative contribution of different colored noises in capturing the whispered speech spectrum could discriminate whispered V and UV phonemes well.

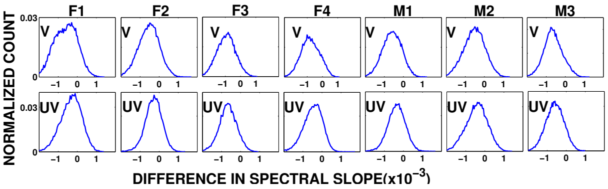

Figure 2: *Histogram of the difference in the spectral slopes computed from the original and the reconstructed (using NMF) V spectra (top row) and UV spectra (bottom row) for seven subjects.*

3. Proposed V/UV classification Method

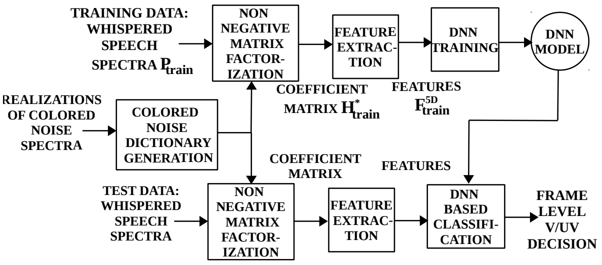
TRAINING DATA: WHISPERED SPEECH SPECTRA P_{train}
NON NEGATIVE MATRIX FACTOR-IZATION
FEATURE EXTRAC-TION
DNN TRAINING
DNN MODEL
COEFFICIENT MATRIX H^*_{train}
FEATURES F^{5D}_{train}
REALIZATIONS OF COLORED NOISE SPECTRA
COLORED NOISE DICTIONARY GENERATION
COEFFICIENT MATRIX
FEATURES
TEST DATA: WHISPERED SPEECH SPECTRA
NON NEGATIVE MATRIX FACTOR-IZATION
FEATURE EXTRAC-TION
DNN BASED CLASSIFI-CATION
FRAME LEVEL V/UV DECISION

Figure 3: *Block diagram highlighting the steps involved in the proposed automatic V/UV classification from whispered speech.*

The block diagram of the proposed DNN based classifier is shown in Fig. 3. We generate W_N by randomly selecting equal number of spectra from each of the five colored noises. We use a set of whispered utterances for training the DNN. With W_N and the spectra computed from the training dataset, P_{train}, we compute the coefficient matrix H^*_{train} by NMF, using Eq. 1. A five dimensional feature F^{5D}_{train} is then computed from H^*_{train} as follows:

$$F^{5D}_{train}[j,p] = \sum_{i=\left(\frac{r}{5}\right)(j-1)+1}^{\left(\frac{r}{5}\right)j} \log\left(H^*_{train}[i,p]\right), \quad j = 1 \ldots 5, \qquad (2)$$

where, $F^{5D}_{train}[j,p]$[2] corresponds to the j^{th} feature in the p-th frame. From Eq. 2, we see that the F^{5D}_{train} captures the strength of the contribution of the CNBVs corresponding to each colored noise. We then train a DNN using these features. Given a test whispered spectrum, we follow the same procedure to compute the features, as depicted in Fig. 3, and perform the V/UV classification using the trained DNN (5D-DNN scheme). The parameters such as the rank r for the NMF and those of the DNN are optimized over a validation dataset. We now explain the dataset used in our experiments.

4. Dataset

For our experiments, we collected whispered speech data from four female (F1, F2, F3 and F4) and three male subjects (M1,

[2] Using the r dimensional features directly from H^*_{train} did not improve the classification accuracy.

M2 and M3), whose native language is Kannada. The subjects are proficient in speaking English. The average age of the subjects was $21(\pm 1.528)$ years. We chose the 460 phonetically balanced sentences from the MOCHA-TIMIT database [14], as the stimuli. The stimuli were presented to the subject and the subject was instructed to whisper each of these 460 sentences. The recordings were carried out in an anechoic chamber using a Sennheizer $e822S$ microphone at a sampling frequency of 16kHz. Since whispered speech is typically of low intensity, there is a need to perform the sound pressure level (SPL) calibration [15]. Therefore, we collected the SPL readings, periodically, during the course of the recordings using a TES-1350A sound level meter. Each recording was listened to and the utterances with errors such as mis-pronunciations, were discarded. For the seven subjects, we obtained 435, 436, 409, 455, 445, 444 and 444 sentences, with a total duration of 153.727 minutes. The average duration of each sentence is $2.861(\pm 0.728)$ seconds.

5. Experiments

5.1. Data Preparation

In order to perform frame-level V/UV classification, we require the ground truth locations of the V and UV segments. Therefore, with the collected data, we perform a forced alignment using a Gaussian mixture model-Hidden Markov model (GMM-HMM) setup, using the Kaldi toolkit [16], with three phonemes – V, UV and silence. The forced aligned boundaries are manually corrected in case of any errors. The V and UV boundaries are then obtained from the corresponding forced aligned V phonemes and UV phonemes, respectively. Since, the goal of the frame-level V/UV classification is to help in obtaining a better pitch contour for whisper to neutral speech conversion, we include ‘silence’ into the UV category.

5.2. Experimental Setup

The collected data shows a class imbalance, with $1.911\,(\pm 0.513)$s of V phonemes and $0.930\,(\pm 0.383)$s of UV phonemes (on average) per utterance. Thus, a scheme that results in a high average, yet a skewed performance in V/UV classification, may not necessarily yield perceptually relevant voicing decisions to reconstruct neutral speech from whispered speech. Hence, we aim to achieve *a balanced or an equal classification performance for both the classes. Therefore, we report two additional numbers– the individual classification accuracies for the two classes, namely, V accuracy and UV accuracy.*

We perform two sets of experiments using the subject-wise and ‘leave-one-subject-out’ setups. For the subject-wise experiments, we use a four-fold setup, where in each fold a randomly picked set of 100 sentences (from one subject) is split, in the ratio $4:1$, to create the training set[3] and the validation set, respectively. A non-overlapping set of 300 sentences (from the same subject) is chosen as the test data. Hence, for each subject we obtain four sets of classification accuracies (for V and UV, individually and averaged) from four folds. For the ‘leave-one-subject-out’ framework, we perform seven experiments, each, using data from six subjects for training and that from the ‘left-out’ subject for testing. We use the test sets from the subject-wise four folds in each of the seven experiments. Therefore, we obtain a total of 28 sets of classification accuracies from 7 subjects $\times$ 4 folds.

5.3. Baseline Schemes

We consider two baseline features to compare the performance of the proposed 5D features. As one baseline scheme, we consider a DNN classifier with the 13-dimensional static Mel Frequency Cepstral Coefficients (MFCC) as features (MFCC-DNN scheme), as they are known to effectively capture the spectral shape of speech. We test the robustness of the constructed CND against a data driven dictionary (DD-DNN scheme). Using NMF[4], we find the dictionary by solving the Euclidean norm ($\|.\|_2$) minimization problem in the training set, $W^* = \arg\min_{W\geq 0, H\geq 0} \|P_{train} - W \times H\|_2$, where, W is a randomly initialized basis matrix, H is the corresponding coefficient matrix and the W^* is the optimized dictionary. We learn separate dictionaries for V, W_V^*, and UV, W_{UV}^*, each of rank r_{dd}. We then concatenate these two dictionaries to obtain the data driven dictionary, $W_{DD} = [W_V^*, W_{UV}^*]$ of dimension $N_F \times 2r_{dd}$. We then compute a two-dimensional feature, F_{train}^{DD}, in a manner similar to Eq. 2.

5.4. Parameters

For our experiments, we compute the spectrogram with a window length, N_w=160 samples, corresponding to 10ms with a shift of N_{sh}=160 samples using FFT bin size of 512. Since, the slow rise and fall of energy from 0 to 1kHz and from 7 to 8kHz, respectively, could affect the estimation of H^* using the CN-BVs, we consider the frequency range 1-7kHz with N_F=193. MFCCs are computed using the same values of N_w and N_{sh} using the Kaldi toolkit. As mentioned in Section 5.2, the choice of the optimal parameters is based on the balance in performance achieved for both the classes, on the validation dataset. From five choices of r, namely, $5, 10, 15, 20$ and 25, we find r=25 as the optimal choice. Similarly, from five choices of r_{dd} being $2, 4, 5, 10$ and 12, r_{dd}=5 is found to be the optimal choice. We implement NMF using the NMFlib package [17].

To understand the nature of voicing information carried by the 5D features and MFCCs, we perform experiments combining the two sets of features (Combined-DNN scheme). In all the schemes, we use a three layer DNN with 64 hidden neurons in each layer. Optimization is done using Adam [18], with a batch size of 5 and binary cross-entropy as the loss function. Based on the performance on the validation dataset, we choose sigmoid activation function for the output layer in all schemes, the ‘relu’ activation function for the hidden layers of the 5D-DNN scheme and ‘tanh’ for the rest. The DNNs are implemented using Keras [19] and Theano [20] libraries.

6. Results and Discussion

6.1. Color noise dictionary Vs data driven dictionary

For better illustration, we choose the same value for r and r_{dd} as 25, to construct W_N, W_V^* and W_{UV}^*. Fig. 4 shows these dictionaries computed from the training data for the second fold of the subject F2. As seen from the figure, the CND contains equal number of realizations of the five different noises with the spectral slope decreasing from left (Violet) to right (Brown). From Fig. 4(b) we see that W_V^* picks up lower frequencies, less than

[3] The training feature set is ensured to have an equal number of V and UV frames to have a balance between the two classes while training.

[4] We use NMF without sparsity constraints on W, since we do not expect the dictionary for the noise-like whispered speech to be sparse.

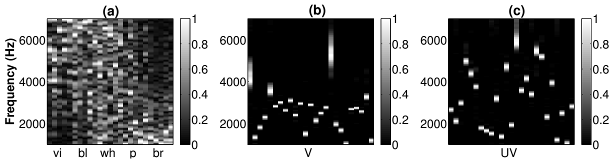

Figure 4: *Dictionaries from one fold of subject F2: (a) W_N constructed using Violet (vi), Blue (bl), White (wh), Pink (p) and Brown (br) noises; (b) Data driven dictionary for V, W_V^* and (c) UV, W_{UV}^*, phonemes.*

3000Hz, corresponding to the typical range of the formant frequencies observed in female whispered speech [21]. Interestingly, we see that a few higher frequencies are also dominant in the V dictionary. This could be due to the fact that whispered speech is characterized by high energy in the higher frequencies [13]. We also observe that W_{UV}^* has more contribution from the higher frequencies, typically above 3500Hz.

6.2. Subject-wise Experimental Results

Averaged across V/UV, we obtain an accuracy of 65.60% (±11.37), 72.26% (±5.60), 77.21% (±6.40), 75.98% (±4.78) for DD-DNN, 5D-DNN, MFCC-DNN and Combined-DNN schemes, respectively. The average accuracy of the DD-DNN scheme is 74.23%(±10.62) and 56.98%(±12.12) for V and UV, respectively. It is clear that the DD-DNN scheme predicts most frames as V than UV. Interestingly, we see that the 5D-DNN shows an improvement of 1.27% and 12.03% (on average) in the V and UV accuracies compared to the DD-DNN scheme. This reveals the potential for a CND to represent, better, the whispered V and UV phonemes compared to a data driven dictionary.

Fig. 5 shows the average V and UV classification accuracies across four folds of the seven subjects for the MFCC-DNN, 5D-DNN and the Combined-DNN schemes. From the figure, we see that the MFCC-DNN scheme predicts most frames as UV (on average) with the drop in the accuracy of UV to V being 11.61% (relative). We see that the 5D-DNN shows a balanced performance (on average) for both V and UV except for subjects F1 and F2. Interestingly, we see that the Combined-DNN scheme bridges the fall in the accuracy from UV to V in the MFCC-DNN scheme of 11.61% (relative) to 3.94% (relative). Also, averaged across V and UV, we obtain comparable accuracies of 77.22%(±4.78) and 76.41%(±4.58) for MFCC-DNN and Combined-DNN, respectively. The Combined-DNN scheme shows an improvement of 4% (relative) in the V accuracy compared to that of MFCC-DNN and 12.94% (relative) in the UV accuracy compared to that of 5D-DNN. This indicates that the information regarding the voicing decisions captured by the 5D features is complementary to that by MFCCs.

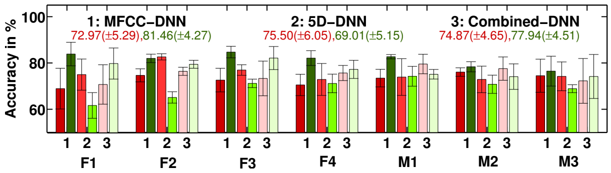

Figure 5: *Average classification accuracies for the 7 subjects across 4 folds, for V, and UV in shades of red and green, repectively, for (1) MFCC-DNN, (2) 5D-DNN and (3) Combined-DNN schemes. Error bars indicate the standard deviation. The average accuracy (standard deviation) across all subjects and all folds is indicated in red text for V and green text for UV for the three schemes.*

6.3. ‘Leave-one-subject-out’ Experimental Results

Table 1: *The average (standard deviation) of the V and UV accuracies, in %, across all folds of the seven subjects*

Schemes	V	UV	Average V/UV
DD-DNN	77.39 (11.82)	53.11 (10.91)	65.25 (11.37)
5D-DNN	76.70 (8.02)	67.39 (4.50)	72.04 (6.26)
MFCC-DNN	73.63 (5.55)	78.51 (4.37)	76.06 (4.87)
Combined-DNN	73.81 (8.31)	74.78 (10.91)	74.29 (6.34)

Table. 1 provides the accuracy within and across V and UV phonemes, averaged over all the folds of the seven subjects. Similar to the observations in Section 6.2, we see that the DD-DNN scheme, shows a skewed performance with the V accuracy being higher than that of UV by 31.37% (relative). We see that this gap is reduced by 5D-DNN to 12.14% (relative). This, together with an increase in the average V/UV accuracy indicates that, the generic CND represents the spectra from an ‘unseen’ subject, better than a subjects specific dictionary learnt from the data. In the MFCC-DNN scheme, the drop in the accuracy from UV to V is 6.63% (relative), a reduction in the gap compared to 5D-DNN. It could be that the nature of the ‘colors’ of V and UV phonemes, is subject dependent. Interestingly, we find that compared to the subject-wise setup, the objective function value obtained while optimizing H^* (using Eq. 1), in the ‘leave-one-subject-out’ setup, turns out to be 2%, 2.21% higher for F1 & F2 and 7.42%, 3.18%, 5.09%, 1.66% and 1.81% lower for the five other subjects (on average). The increase seen for subjects F1 and F2 indicates that the ‘coloring’ of the spectra for the two subjects is different from that of the other subjects and, hence, is poorly represented in the ‘leave-one-subject-out’ setup compared to the subject-wise setup. Finally, we find that the Combined-DNN scheme, yields the most balanced performance, with an average V/UV accuracy comparable to that of MFCC-DNN, in addition to the fall of accuracy from UV to V being only 1.30%, the least among all schemes,

7. Conclusion

We perform a frame-level classification of whispered V and UV phonemes, by exploiting the noise-like nature of whispered speech. Interestingly, we see that the CND represents the whispered speech spectra, better, compared to a data driven dictionary. Experiments both, within and across subjects, reveal that the contribution of each CNBV, varies from V to UV, confirming that the ‘color’ of the V and UV phonemes is, indeed, different. We also find that the features extracted from the CND carry complementary information regarding voicing decisions, compared to the MFCCs. This makes the scheme, trained on the combination of the two features, exhibit a more balanced performance across the two classes. Further analysis is required to understand the broad class phoneme specific coloring in whispered speech and their dominant CNBVs. Investigating the hypothesis of a subject dependent ‘coloring’ of the V and UV phonemes using articulatory data, is a part of our future work.

8. Acknowledgment

We thank our seven subjects and express our gratitude to Kausthubha N.K. for the manual annotation of the data.

9. References

[1] H. R. Sharifzadeh, I. V. McLoughlin, and F. Ahmadi, "Reconstruction of normal sounding speech for laryngectomy patients through a modified CELP codec," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 10, pp. 2448–2458, 2010.

[2] V. C. Tartter, "Whats in a whisper?" *The Journal of the Acoustical Society of America*, vol. 86, no. 5, pp. 1678–1683, 1989.

[3] R. W. Morris and M. A. Clements, "Reconstruction of speech from whispers," *Medical Engineering & Physics*, vol. 24, no. 7, pp. 515–520, 2002.

[4] F. Ahmadi, I. V. McLoughlin, and H. R. Sharifzadeh, "Analysis-by-synthesis method for whisper-speech reconstruction," in *Asia Pacific Conference on Circuits and Systems, APCCAS*. IEEE, 2008, pp. 1280–1283.

[5] M. Janke, M. Wand, T. Heistermann, T. Schultz, and K. Prahallad, "Fundamental frequency generation for whisper-to-audible speech conversion," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 2579–2583.

[6] G. N. Meenakshi and P. K. Ghosh, "A discriminative analysis within and across voiced and unvoiced consonants in neutral and whispered speech in multiple indian languages," in *INTERSPEECH*, 2015, pp. 781–785.

[7] T. Toda and K. Shikano, "NAM-to-speech conversion with Gaussian mixture models," in *INTERSPEECH*, 2005, pp. 1957–1960.

[8] I. V. Mcloughlin, H. R. Sharifzadeh, S. L. Tan, J. Li, and Y. Song, "Reconstruction of phonated speech from whispers using formant-derived plausible pitch modulation," *ACM Transactions on Accessible Computing (TACCESS)*, vol. 6, no. 4, p. 12, 2015.

[9] M. F. Schwartz, "Identification of speaker sex from isolated, voiceless fricatives," *The Journal of the Acoustical Society of America*, vol. 43, no. 5, pp. 1178–1179, 1968.

[10] A. Löfqvist and B. Mandersson, "Long-time average spectrum of speech and voice analysis," *Folia Phoniatrica et Logopaedica*, vol. 39, no. 5, pp. 221–229, 1987.

[11] S. Ghaffarzadegan, H. Boril, and J. H. Hansen, "UT-VOCAL EFFORT II: Analysis and constrained-lexicon recognition of whispered speech," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 2544–2548.

[12] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001, pp. 556–562.

[13] C. Zhang and J. H. Hansen, "Analysis and classification of speech mode: whispered through shouted." in *INTERSPEECH*, vol. 7, 2007, pp. 2289–2292.

[14] A. Wrench, "MOCHA-TIMIT," *Department of Speech and Language Sciences, Queen Margaret University College, Edinburgh, speech database*, 1999.

[15] C. Zhang and J. H. L. Hansen, "Advancements in whisper-island detection within normally phonated audio streams," in *Proc. Interspeech*, 2009, pp. 860–863.

[16] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech Recognition Toolkit," in *IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, 2011.

[17] G. Grindlay, "NMFLib - efficient matlab library implementing a number of common nmf variants," *URL-http://www.ee.columbia.edu/ grindlay/code.html*, 2010.

[18] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

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

[20] T. T. D. Team, R. Al-Rfou, G. Alain, A. Almahairi, C. Angermueller, D. Bahdanau, N. Ballas, F. Bastien, J. Bayer, A. Belikov *et al.*, "Theano: A python framework for fast computation of mathematical expressions," *arXiv preprint arXiv:1605.02688*, 2016.

[21] H. R. Sharifzadeh, I. V. McLoughlin, and M. J. Russell, "A comprehensive vowel space for whispered speech," *Journal of voice*, vol. 26, no. 2, pp. e49–e56, 2012.