Analysis of whispered speech and its conversion to neutral speech

G. Nisha Meenakshi

SPIRE Lab, Electrical Engineering, Indian Institute of Science, Bangalore-560012.

nishag@iisc.ac.in

1. Research Problem Statement

Whispering is an indispensable form of communication that emerges in private conversations as well as in pathological situations [1]. In conditions such as partial or total laryngectomy, spasmodic dysphonia etc, alaryngeal speech such as esophageal, tracheo-esophageal speech and hoarse whispered speech are common [2, 3]. Whispered speech is primarily characterized by the lack of vocal fold vibrations, and, hence, pitch [4]. In addition to lack of voicing, whispered speech is also characterized by formant shifts [5], that are typically due to the exaggerated articulatory movements exhibited while whispering. In recent times, applications such as voice activity detection [6, 7], speaker identification and verification [8, 9, 10] and speech recognition [11, 12, 13] have been extended to whispered speech as well. Several efforts have also been undertaken to convert the less intelligible whispered speech into a more natural sounding neutral speech [14, 15, 16, 17]. While there have been a few works in the literature, a thorough understanding of the acoustic and articulatory characteristics of whispering remain unclear. Hence, the aim of the thesis is two-fold, 1) to analyze the different characteristics of whispered speech using both speech and articulatory data, 2) to perform whispered speech to neutral speech conversion using the state-of-the art modeling techniques.

2. Problems addressed and key findings

2.1. Analysis of whispered speech

2.1.1. Analysis employing speech data

It is known that pitch, that results from vocal fold vibrations carries the speaker's gender related information in the case of neutral speech. It is now of interest to understand how whispered speech, which is typically unvoiced, encodes the information about the speaker's gender [18]. Automatic gender classification using mel-frequency cepstral coefficients (MFCC) and a support vector machine (SVM) classifier reveals that whispered speech results in a classification accuracy of 89.43%, while neutral speech results in 93.84%. This indicates that although the gender specific information is captured better by neutral speech, whispered speech is not completely devoid of it. Further analysis also shows that 1) the first MFCC is rich in gender specific information compared to the others and 2) the dynamics of MFCCs do not carry additional gender related information compared to their static counterparts, in both neutral and whispered speech.

There are several challenges involved in developing a whispered to neutral conversion system. Apart from modifying the spectrum of whispered speech, there is a need to estimate and incorporate pitch. It is known that pitch exists only in the case of voiced (V) phonemes, whose production involves vocal fold vibrations, unlike unvoiced (UV) phonemes, where the vocal folds remain open. Therefore in order to incorporate pitch, we first require the boundaries of V and UV phonemes from whispered speech. It is now of interest to first understand if humans perceive the difference between the whispered V and UV phonemes. Therefore, we investigate if the discrimination between the whispered V-UV consonants is still preserved using seven V-UV consonant pairs, in the form of vowel consonant-vowel (VCV) stimuli, from six Indian languages [19]. Subjective analysis reveals that even with the lack of voicing in whispered speech, voiced consonants do not completely lose their identity, although, on average, the discrimination of V-UV pairs in whispered speech is lower compared to that in neutral speech. It is also found that the variation of the acoustics from neutral to whispered speech is consonant specific. Based on objective analysis, it is seen that the acoustic properties of the stop consonants are preserved more than the affricates in whispered speech. Specifically, it is observed that while the Velar stop consonant (/g/) is most affected, the Palato-alveolar sibilant (/ʒ/) is least affected. Since this work indicates that V-UV phonemes can be discriminated from whispered speech, a potential bases to represent the whispered V and UV spectra is further explored.

Exploiting the noise-like nature of whispered speech, a hypothesis, that a whispered speech spectrum can be represented as a linear combination of a set of colored noise spectra, is made [20]. Five colored noises with increasing spectral slope f and f^2, decreasing spectral slope $1/f$ and $1/f^2$ and white Gaussian noise are considered. A colored noise dictionary (CND) using spectra from each of the five noises is constructed and non-negative matrix factorization (NMF) [21] is employed to estimate the contribution from each of these colored noise basis vectors (CNBV). In order to find if these colored noises could represent the whispered V and UV phonemes well, we perform an automatic classification of V-UV with features computed using NMF on the CND, employing a Deep Neural Network (DNN) as the classifier. We also compare the performance of the CND with 1) features computed from a data driven dictionary and 2) MFCCs. Averaged across V/UV, we obtain an accuracy of $65.60\%(\pm 11.37)$, $72.26\%(\pm 5.60)$ and $77.21\%(\pm 6.40)$ for the data driven dictionary, CND and MFCC schemes, respectively. It is observed that the CND represents the whispered speech spectra, better, compared to a data driven dictionary and performs comparable to that of MFCCs. The observation that the contribution of each CNBV, varies from V to UV, confirms that the 'color' of the whispered V and UV phonemes is, indeed, different. These studies demonstrate that although unvoiced, the spectrum of whispered speech carries voicing cues. These voicing cues could be attributed to the exaggerated movements of the articulators while whispering [22, 23, 24, 25].

2.1.2. Analysis employing articulatory data

Since it is known that whispered and neutral speech have different characteristics in both acoustics and articulation, a comparison of the accuracy with which the articulation can be recovered

from the acoustics is done from both types of speech, individually [26]. The 460 phonetically balanced English sentences from the MOCHA-TIMIT database [27] are used as stimuli for recording both the audio and articulatory data using electromagnetic articulography (EMA) [28] from four subjects, along the mid-sagittal plane from six sensors, namely, upper lip (UL), lower lip (LL), jaw (J), tongue tip (TT), tongue body (TB) and tongue dorsum (TD). Twelve dimensional articulatory features are computed (from the two axes of the plane, eg. UL_x, UL_z). Acoustic-to-articulatory inversion (AAI) performed with these twelve articulatory features using a DNN reveals that the performance drops significantly only for J_x, TT_z and TB_z in the case of whispered speech compared to neutral speech. This suggests that although whispered speech lacks voicing and is less intelligible compared to neutral speech, the information about the articulatory movements is still encoded in the spectral characteristics of whispered speech. Experiments with models trained on neutral speech being tested by whispered speech and vice-versa demonstrate that acoustic-to-articulatory mapping of whispered speech is different from that of the neutral speech. Further analysis also reveals that the dynamics of the articulatory movements while whispering is smoother than that of neutral speech. These observations trigger a bigger question– 'how are the articulatory movements in whispered speech, related to those in neutral speech?'

In order to investigate how the neutral speech articulatory trajectories (NATs) are related to the whispered speech articulatory trajectories (WATs), three candidate transformation functions (TF), namely, an affine function with a diagonal matrix ($\mathcal{A}_d$), a full matrix ($\mathcal{A}_f$) and a DNN based nonlinear function are considered [29]. While $\mathcal{A}_d$ reconstructs one NAT from the corresponding WAT, $\mathcal{A}_f$ and DNN reconstruct each NAT from all WATs. For these experiments, EMA recordings from six subjects are considered with the 460 phonetically balanced utterances from the MOCHA-TIMIT database as the stimuli. The TF which results in the least dynamic time warped (DTW) [30] distance between the transformed WATs and the original NATs turns out to be the affine function with a full matrix $\mathcal{A}_f$, at utterance level and across broad class phoneme categories. This indicates that the exaggerated articulatory movements in whispered speech need not results in a highly nonlinear transformation between the WATs and NATs. Further investigation shows that the sensors placed on the tongue, show an increase in, both, stability and precision in their movements, while whispering. This indicates that controlling the articulation of tongue is vital to improve the intelligibility of whispered speech, compared to the other articulators considered in the study.

2.2. Whispered to neutral speech conversion

Detection of whispered speech from a stream of audio is an important preprocessing module that precedes the conversion step. Thus, whisper activity detection (WAD) is done to segment whispered speech, given a noisy recording of whispered speech [31]. Unlike voice activity detection of neutral speech from a noisy recording, WAD is even more challenging due to the noise-like nature of whispered speech. A feature based on the long term variation of the logarithm of the the sub-band signal's energy profile is proposed to detect whispered speech. Experiments are performed with eight different noises at four different signal-to-noise ratios. The sub-bands that are vital to discriminate noise from noisy whispered speech turn out to be noise specific. It is also found that the proposed feature outperforms four other baseline schemes even at -5dB SNR.

In order to reconstruct neutral speech from whispered speech, methods that directly modify the formants and compute pitch from formants have been proposed [14, 15]. In addition to such methods, voice conversion based techniques using Gaussian mixture models [17] and DNNs [16] have also been proposed. Since these techniques do not exploit the time varying structure in speech, a bi-directional LSTM (BLSTM) [32, 33] based whispered to neutral speech conversion is proposed [34]. In this work, the STRAIGHT speech synthesizer is used. Four BLSTMs are trained to estimate the neutral mel cepstral coefficients, pitch, periodicity level and the V-UV class from whispered mel cepstral coefficients. Subject specific experiments using data from six subjects reveal that the BLSTM estimates both the spectral and excitation features, better than its DNN counterpart. It is also observed that the pitch prediction using the two models is comparable. In a listening test the proposed BLSTM based scheme is chosen $\sim 27\%$ more often than a DNN based baseline scheme. This subjective evaluation confirms that the proposed BLSTM based whispered to neutral conversion scheme results in a more natural sounding reconstructed speech. Further analysis following feedback from the participants of the listening test reveals that the improvement in the naturalness of the proposed scheme could be due to the temporally smoother feature trajectories predicted by the BLSTM.

3. Major contributions of the research

The summary of the major contributions of the research is as follows:

- An improved understanding about how whispered speech maintains its degree of intelligibility even with the lack of voicing. This is done by the analysis of how the whispered speech spectrum encodes speaker gender related information, detailed investigation of human perception of whispered V-UV phonemes and proposal of bases for V-UV phonemes motivated by the noise-like nature of whispered speech.
- Detailed understanding of how articulation varies across neutral and whispered speech with regard to differences in the performance of AAI, investigation of the transformation function that relates neutral and whispered articulation and analysis of exaggerated articulatory movements in whispered speech using EMA.
- Whisper activity detection under noisy condition.
- Whispered to neutral speech conversion using bi-directional LSTMS.

3.1. Problems to be addressed

The problems to be addressed include understanding 1) how the whispered spectrum must be parametrized in order to aid both effective representation and mapping of whispered speech into neutral speech, 2) how pitch is encoded in whispered articulatory movements and 3) how stress manifests itself in whispered speech.

4. Acknowledgements

The author would like to thank Dr. Prasanta Kumar Ghosh (Advisor) for his guidance, the subjects for the recordings, Mr. Aravind Illa and Mrs. Kausthubha N.K for the manual processing of the data.

5. 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] W. Wszolek, M. Modrzejewski, and M. Przysikeny, "Acoustic analysis of esophageal speech in patients after total laryngectomy," *Archives of Acoustics*, vol. 32, no. 4 (S), pp. 151–158, 2014.

[3] A. G. Gilchrist, "Rehabilitation after laryngectomy," *Acta Oto-Laryngologica*, vol. 75, no. 2-6, pp. 511–518, 1973. [Online]. Available: https://doi.org/10.3109/00016487309139782

[4] V. C. Tartter, "What's in a whisper?" *J. Acoust. Soc. Amer.*, vol. 86, no. 5, pp. 1678–1683, 1989.

[5] 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.

[6] C. Zhang and J. H. Hansen, "Whisper-island detection based on unsupervised segmentation with entropy-based speech feature processing," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 883–894, 2011.

[7] M. Sarria-Paja and T. H. Falk, "Whispered speech detection in noise using auditory-inspired modulation spectrum features," *IEEE Signal Processing Lett.*, vol. 20, no. 8, pp. 783–786, 2013.

[8] X. Fan and J. H. Hansen, "Speaker identification within whispered speech audio streams," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 5, pp. 1408–1421, 2011.

[9] X. Fan and J. H. Hansen, "Speaker identification for whispered speech based on frequency warping and score competition." in *Proc. ISCA Interspeech*, 2008, pp. 1313–1316.

[10] X. Fan and J. H. Hansen, "Speaker identification for whispered speech using modified temporal patterns and MFCCs." in *Proc. ISCA Interspeech*, 2009, pp. 896–899.

[11] T. Ito, K. Takeda, and F. Itakura, "Analysis and recognition of whispered speech," *Speech Commun.*, vol. 45, no. 2, pp. 139–152, 2005.

[12] T. Itoh, K. Takeda, and F. Itakura, "Acoustic analysis and recognition of whispered speech," in *Proc. IEEE ICASSP*, 2002, pp. 389–392.

[13] S.-C. S. Jou, T. Schultz, and A. Waibel, "Whispery Speech Recognition using Adapted Articulatory Features." in *Proc. IEEE ICASSP*, 2005, pp. 1009–1012.

[14] R. W. Morris and M. A. Clements, "Reconstruction of speech from whispers," *Med. Eng. Phys.*, vol. 24, no. 7, pp. 515–520, 2002.

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

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

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

[18] G. N. Meenakshi and P. K. Ghosh, "Automatic gender classification using the mel frequency cepstrum of neutral and whispered speech: A comparative study," in *Communications (NCC), 2015 Twenty First National Conference on*. IEEE, 2015, pp. 1–6.

[19] 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.

[20] G. N. Meenakshi and P. K. Ghosh, "A robust voiced/unvoiced phoneme classification from whispered speech using the color of whispered phonemes and deep neural network," in *Proc. Interspeech 2017*, 2017, pp. 503–507. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-1388

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

[22] H. Yoshioka, "The role of tongue articulation for /s/ and /z/ production in whispered speech," in *Proc. Acoustics*, 2008, pp. 2335–2338.

[23] M. J. Osfar, "Articulation of whispered alveolar consonants," Master's thesis, University of Illinois at Urbana-Champaign, 2011.

[24] M. F. Schwartz, "Bilabial Closure Durations for /p/, /b/, and /m/ in Voiced and Whispered Vowel Environments," *J. Acoust. Soc. Amer.*, vol. 51, pp. 2025–2029, 1972.

[25] M. Parnell, J. D. Amerman, and G. B. Wells, "Closure and constriction duration for alveolar consonants during voiced and whispered speaking conditions," *J. Acoust. Soc. Amer.*, vol. 61, pp. 612–613, 1977.

[26] A. Illa, N. M. G, and P. K. Ghosh, "A comparative study of acoustic-to-articulatory inversion for neutral and whispered speech," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 5075–5079.

[27] A. Wrench, "MOCHA-TIMIT," Department of Speech and Language Sciences, Queen Margaret University College, Edinburgh, speech database, 1999. [Online]. Available: http://sls.qmuc.ac.uk

[28] P. W. Schnle, K. Grbe, P. Wenig, J. Hhne, J. Schrader, and B. Conrad, "Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract," *Brain and Language*, vol. 31, no. 1, pp. 26 – 35, 1987. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0093934X87900587

[29] N. M. G. and P. K. Ghosh, "Reconstruction of articulatory movements during neutral speech from those during whispered speech," *The Journal of the Acoustical Society of America*, vol. 143, no. 6, pp. 3352–3364, 2018. [Online]. Available: https://doi.org/10.1121/1.5039750

[30] M. Müller, "Dynamic time warping," *Information retrieval for music and motion*, pp. 69–84, 2007.

[31] G. N. Meenakshi and P. K. Ghosh, "Robust whisper activity detection using long-term log energy variation of sub-band signal," *IEEE Signal Processing Letters*, vol. 22, no. 11, pp. 1859–1863, 2015.

[32] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, Nov 1997.

[33] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[34] G. N. Meenakshi and P. K. Ghosh, "Whispered speech to neutral speech conversion using bidirectional lstms," in *INTERSPEECH*, 2018.