Interspeech 2018
2-6 September 2018, Hyderabad

Whispered speech to neutral speech conversion using bidirectional LSTMs

G. Nisha Meenakshi, Prasanta Kumar Ghosh

Electrical Engineering, Indian Institute of Science, Bangalore- 560012, Karnataka, India
nishag@iisc.ac.in, prasantg@iisc.ac.in

Abstract

We propose a bidirectional long short-term memory (BLSTM) based whispered speech to neutral speech conversion system that employs the STRAIGHT speech synthesizer. We use a BLSTM to map the spectral features of whispered speech to those of neutral speech. Three other BLSTMs are employed to predict the pitch, periodicity levels and the voiced/unvoiced phoneme decisions from the spectral features of whispered speech. We use objective measures to quantify the quality of the predicted spectral features and excitation parameters, using data recorded from six subjects, in a four fold setup. We find that the temporal smoothness of the spectral features predicted using the proposed BLSTM based system is statistically more compared to that predicted using deep neural network based baseline schemes. We also observe that while the performance of the proposed system is comparable to the baseline scheme for pitch prediction, it is superior in terms of classifying voicing decisions and predicting periodicity levels. From subjective evaluation via listening test, we find that the proposed method is chosen as the best performing scheme 26.61% (absolute) more often than the best baseline scheme. This reveals that the proposed method yields a more natural sounding neutral speech from whispered speech.

Index Terms: Whispered speech, LSTM, STRAIGHT.

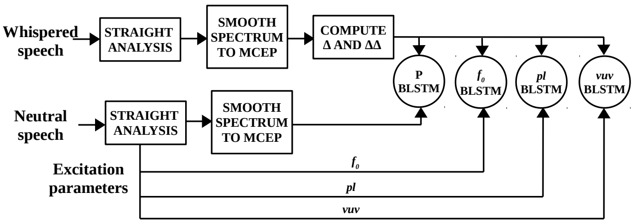


Figure 1: *Illustration of the steps for training the BLSTMs used to convert whispered to neutral speech.*

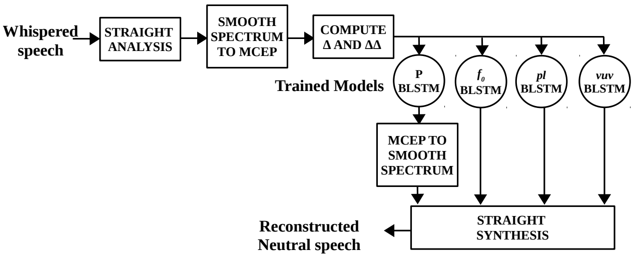


Figure 2: *Illustration of the steps for the BLSTM based whispered to neutral speech conversion system.*

1. Introduction

Whispered speech is a natural mode of speech production, typically produced in pathological cases, such as laryngectomy [1], as well as in private conversations. Whispered speech lacks pitch due to the absence of vocal fold vibrations during its production [2]. Several attempts have been made in the past to convert the voiceless whispered speech into neutral speech, e.g., the silent speech interfaces [3]. There exist several differences between the spectra of whispered speech and neutral speech. Several spectral characteristics of whispered speech have been reported in the literature including, the shift of formants [4], gender based differences in this formant shift [5, 6], differences in the spectra of neutral and whispered voiced and unvoiced phonemes [7]. Therefore, whispered to neutral speech conversion systems need to modify the spectrum of whispered speech, in addition to estimating and incorporating pitch using appropriate voicing decisions to reconstruct neutral speech.

One class of methods to reconstruct neutral speech from whispered speech involves the modification of the whispered speech spectrum by shifting formants (using empirically computed shift) [1, 8] followed by the incorporation pitch estimated from formants [9]. The mixed excitation linear prediction [10] and code excited linear prediction based vocoders [8] are used for the synthesis. The other class of methods typically assumes the presence of parallel neutral and whispered speech data to train statistical models to reconstruct neutral speech from whispered speech. In this case, the statistical models are required to predict both the spectral features and excitation parameters of the neutral speech from whispered speech. Specifically, statistical models such as Gaussian mixture models [11] and deep neural networks (DNN) [12] are employed for this purpose. The speech synthesized using these methods is reported to sound less natural [12] owing to the discontinuities introduced by the frame level prediction of the parameters.

In this work, we propose a whispered to neutral speech conversion system that employs bidirectional long short-term memory networks (BLSTM) [13, 14] as the statistical model and STRAIGHT [15] as the speech synthesis module. We employ BLSTMs since they capture the relevant information about the underlying temporal structure in the data, using both the past and the future contexts [16]. We use them to predict the parameterized spectrum, frame level pitch, voicing decisions and aperiodicity levels to synthesize neutral speech from whispered speech. Using both objective and subjective evaluations, we find that the BLSTM based whispered to neutral system yields a more natural sounding speech compared to the DNN based baseline schemes. We begin with the description of the proposed BLSTM based whispered to neutral speech conversion system.

2. Proposed Method

In this work, we employ the STRAIGHT speech synthesis framework for which we require estimates of the smooth

10.21437/Interspeech.2018-1487

spectrum (P), frame level pitch (f_0), voiced-unvoiced (vuv) phoneme decisions and the periodicity levels (pl). Therefore, in the proposed system, parallel training data of whispered speech and neutral speech is used to train four BLSTMs, namely, P-BLSTM, f_0-BLSTM, vuv-BLSTM and pl-BLSTM to predict the parameterized smooth spectrum, pitch, voicing decisions and aperiodicity levels, respectively. Fig. 1 provides a block diagram describing the training procedure for the proposed BLSTM based whispered to neutral speech conversion system. Firstly, we perform STRAIGHT analysis on whispered speech to obtain an estimate of the smooth spectrum P_w. We then compute the m^{th} order mel-cepstral coefficients, $C[0], \ldots, C[m]$, (MCEP) from P_w. Velocity (Δ) and acceleration ($\Delta\Delta$) coefficients are then appended to constitute the feature vectors corresponding to whispered speech. In a similar fashion, we compute the m^{th} order MCEPs from the smooth spectral estimates of the neutral speech as feature vectors to train the P-BLSTM. The three excitation parameters, namely f_0, vuv and pl are obtained from the STRAIGHT analysis of the training data corresponding to neutral speech. Since the duration of the whispered and neutral speech utterances could be different, the neutral and whispered speech is, at first, time aligned prior to training. For this, we employ dynamic time warping (DTW) [17] and obtain the warping path between a mean and standard deviation normalized whispered and neutral MCEPs for each training utterance [1]. The obtained DTW path is used to align the whispered MCEP with not only the neutral MCEP but also the neutral excitation parameters. The four BLSTMs are then trained using the corresponding training data, as depicted in Fig. 1. In the test phase, as shown in Fig. 2, given a test whispered utterance we compute the features MCEP+Δ+$\Delta\Delta$ and use the four trained BLSTM models to predict the neutral MCEP and excitation features. We obtain the smooth spectral estimate from the predicted MCEP and feed both the spectral and the excitation parameters to the STRAIGHT synthesis module to reconstruct the neutral speech. We now describe the dataset used in this work.

3. Dataset

We recorded 460 sentences taken from the MOCHA-TIMIT database [18] from six subjects, three males (M1, M2, M3) and three females (F1, F2, F3). In order to have parallel whispered and neutral speech data, the subjects were asked to speak each sentence in neutral and whispered modes separately. Recordings were done in a sound proof room in five sessions, where in each session a subject would record a set of 100 utterances in neutral and then in whispered speech. Sennheizer e822S microphone was used to record the data and TES-1350A sound level meter was used to perform a sound pressure level calibration [19]. The native language of the six subjects is Kannada, an Indian language. All subjects (with an average age $20.83(\pm 1.60)$ years) are proficient in speaking, reading and writing English. Utterances that contained improper or wrong pronunciations were discarded after manual examination of the recordings. A total of 428, 398, 420, 425, 431, 430 utterances were obtained for F1, F2, F3, M1, M2 and M3, respectively. After silence removal, the total duration turned out to be 125.54 minutes and 124.54 minutes for whispered and neutral speech, respectively, across all subjects.

4. Experiments

The MCEPs of order $m = 25$ and the excitation parameters are computed every 10ms for both whispered and neutral speech. We consider a four fold experimental setup where the data from each subject is divided into training and test sets in a ratio $3 : 1$. From the training dataset, 10% of the data is considered as the validation dataset on which the parameters of the BLSTM are optimized. Specifically, we consider an architecture with an input layer of dimension 78 (26 MCEP + 26 Δ + 26 $\Delta\Delta$), one hidden layer with 'tanh' activation function and one (time-distributed) output layer. The dimension of the output layer is 26 for P-BLSTM and 1 for the other three BLSTMs. For experiments, we consider 32, 64, 128, 256 hidden layers for P-BLSTM and 16, 32, 64 and 128 for the other three BLSTM models. A dropout of 0.1 is used. While mean squared error is used as the loss function for P-BLSTM, f_0-BLSTM and pl-BLSTM, binary cross-entropy is used as the loss function for vuv-BLSTM. In each fold, we choose the model that results in the least error over the validation dataset. Optimization is done using Adam [20]. The implementation of the BLSTM is done using Keras [21] and Theano [22] libraries.

4.1. Baseline Schemes

In order to compare the performance of the proposed BLSTM based whispered to neutral speech conversion system, we choose a DNN based system proposed by Janke *et al.* [12]. In this baseline scheme (B1), the authors use DNNs to predict the neutral MCEP and pitch from whispered MCEP. They use the Mel Log Spectrum Approximation (MLSA) filter [23] method for synthesis. We extend B1 into a second baseline scheme (B2) that uses STRAIGHT for synthesis in place of MLSA filter. Specifically, a neural network architecture to model pitch, proposed in [12], is considered to predict *pl*. The same architecture with a modification to the output layer is employed to predict voicing, i.e., a softmax layer to classify the voiced and unvoiced phonemes. Hence, B2 uses four DNNs to perform whispered to neutral speech conversion. It is to be noted that we compute the MCEP directly from the speech, as proposed in [12], for the two baseline schemes.

4.2. Evaluation Metric

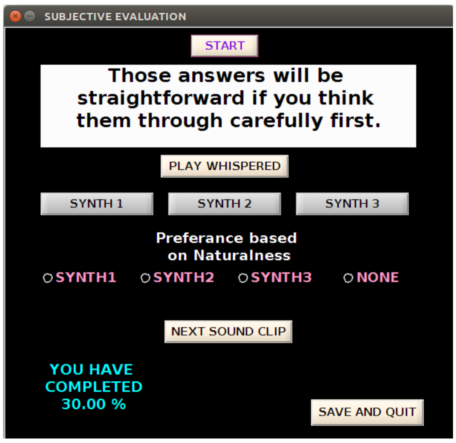
The performance of the proposed BLSTM based whispered to neutral speech conversion system is compared to the two baseline schemes via both objective and subjective evaluations.

4.2.1. Objective Evaluation

We quantify the quality of the predicted MCEP via the objective measures Mel cepstrum distortion (MCD) [24] and the distortion of the 0^{th} cepstral coefficient ($C[0]$ D) [12]. To compute these measures, we first align the predicted MCEP and the MCEP of the corresponding neutral speech via DTW. With regard to the voicing decisions, we report the total error in the prediction of the voiced and unvoiced phonemes. Specifically, let the time aligned (using predicted and original MCEP) voicing decisions of the original and predicted neutral speech of length N, corresponding to the i^{th} test utterance, be indicated by vd_i^o and vd_i^p, respectively. The total error in the prediction of the voiced and unvoiced phonemes for the i^{th} test utterance is defined as, $\frac{1}{N}\sum_{j=1}^{N}\mathcal{I}(vd_i^o[j], vd_i^p[j]) \times 100$ where,

[1] It is to be noted that the mean and standard deviation normalization is done only to find the time alignment.

$\mathcal{I}(vd_i^o[j], vd_i^p[j]) = \begin{cases} 1, & \text{if } vd_i^o[j] = vd_i^p[j] \\ 0, & \text{otherwise} \end{cases}$. The objective measures for pitch and periodicity level for a given test utterance are computed over the voiced frames alone by considering the aligned voicing decisions of the corresponding original neutral speech utterance. Let the pitch contour of the original and predicted neutral speech corresponding to the i^{th} test utterance, considering only the voiced frames be indicated by $f{0}_i^o$ and $f{0}_i^p$ with length $M \leq N$, respectively. We compute percentage change in the predicted and the original pitch (in logarithmic scale) as, $\log\left(\frac{1}{M}\sum_{j=1}^{M}\frac{|f{0}_i^o[j]-f{0}_i^p[j]|}{f{0}_i^o[j]}\times 100\right)$. Similarly, we compute the mean squared error (MSE) in the predicted periodicity level using the time aligned periodicity levels of the original (pl_i^o) and predicted neutral speech (pl_i^p) of length $M \leq N$ as, $\text{MSE}_i = \frac{1}{M}\sum_{j=1}^{M}(pl_i^o[j] - pl_i^p[j])^2$. We report these measures averaged across all the utterances of the test set, in every fold of each of the six subjects. It is to be noted that while the objective measures with regard to the spectral features and pitch are identical for B1 and B2, those with regard to voicing decisions and periodicity level are not applicable to B1.



Those answers will be straightforward if you think them through carefully first.

Preferance based on Naturalness

YOU HAVE COMPLETED 30.00 %

Figure 3: *Graphical user interface used for the subjective evaluation*

4.2.2. Subjective Evaluation

To evaluate the naturalness of the synthesized speech, we performed a perceptual listening test using EX-29 Direct Sound Extreme isolation headphones. From each subject, we choose ninety maximally non-overlapping utterances and evaluate each utterance by three listeners. We considered eighteen listeners, 10 males and 8 females with an average age of 22.33(±1.97) years. The listeners, proficient in reading, writing and speaking English, were not reported to have any hearing disorder. In the graphical user interface (GUI) developed in MATLAB R2014a, shown in Fig. 3, we provided the input whispered speech and the synthesized speech using the two baseline and the proposed schemes (denoted by SYNTH1, SYNTH2 and SYNTH3). The text corresponding to the utterance presented to the listener was also provided in the GUI for the convenience of the listener. The listeners were asked to listen to each of these samples and choose the best synthesized speech based on naturalness. The listeners could also choose none of the three schemes if they deemed fit. Among the 90 utterances, we randomly chose ten utterances and presented them again during the course of the evaluation to check the consistency of the listener. The total duration of the listening test turned out to be 31.89(±7.43) minutes. All the listeners are found to be at least 60% consistent.

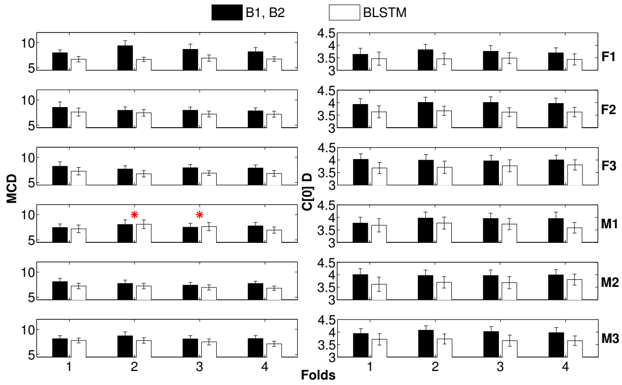


Figure 4: *Average mel cepstral distortion (MCD) and distortion of the 0^{th} cepstral coefficient ($C[0]$ D) across test utterance in each fold of every subject obtained using the proposed and baseline schemes. Error bars indicate standard deviation. The MCD and $C[0]$ D for both B1 and B2 are identical. Hence, only one bar is shown for them.*

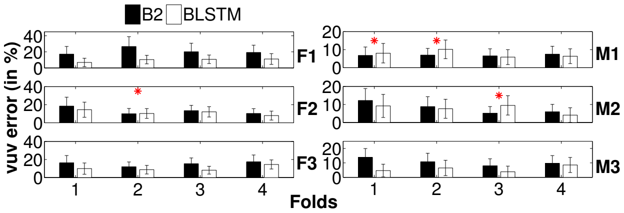


Figure 5: *Percentage error in voiced/unvoiced (vuv) classification for every fold of each subject. Error bars indicate standard deviation.*

5. Results and Discussion

Fig. 4 provides the fold-wise bar plot of the MCD and $C[0]$ D for each of the six subjects for the proposed and the baseline schemes. We find that the average MCD obtained using the proposed BLSTM based whisper to neutral speech conversion system is statistically significantly lower than the baseline schemes (*t-test* p-value $< 1.17e$-12) except in two folds of subject M1 (indicated by red $*$). In case of $C[0]$ D as well, we find that the proposed system exhibits a statistically significantly lower distortion compared to the baseline schemes (p-value $<$ $1.92e$-8)[2]. The average error in the vuv classification (in percentage) across all test utterances in each fold is provided for every subject in Fig. 5. We observe that the classification error is statistically significantly lower (p-value $< 8.60e$-3) using vuv-BLSTM than that by the B2 model except in a few folds (indicated by red $*$ in Fig. 5). This indicates that the performance of a BLSTM based whispered vuv phoneme classifica-

[2] A modified B2 scheme that uses STRAIGHT spectrum based MCEP performed better than B2 but poorer than the BLSTM. Thus, STRAIGHT spectrum based MCEP is a better spectral representation and is better modeled by a BLSTM compared to a DNN.

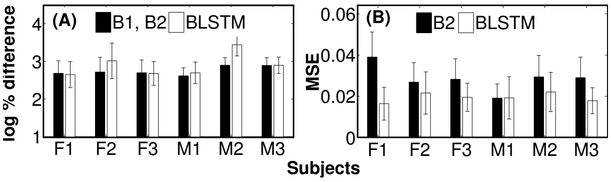


Figure 6: *(A) Percentage change in the predicted and original pitch (in logarithmic scale), (B) Mean squared error in periodicity level prediction, averaged across all folds of each subject. Error bars indicate standard deviation.*

Table 1: *Listening Test Preference Scores*

Methods	B1	B2	BLSTM	None
Preference (in %)	0.43	34.44	61.05	4.07

tion is superior to that by a DNN. Fig. 6 (A) and (B) plot the objective measures to quantify the quality of the pitch and periodicity level prediction, respectively. Using the B2 scheme, we find that in one fold of F1, F3, M3, three folds of F2, M2 and in all folds of M2, the logarithm of the absolute percentage change of pitch is statistically significantly lower (p-value $< 1.08e$-2) compared to that with the f_0-BLSTM. From the figure, we find that the prediction of pitch using a BLSTM is more subject sensitive (the % difference is higher for subjects M2 and F2 compared to the rest) compared to that using a DNN. From Fig. 6 (B) we find that the performance of the pl-BLSTM is superior to that by the baseline scheme. Specifically, we observe that except for two folds of subject M1, in all folds of all subjects the mean squared error in periodicity level prediction is statistically significantly lower (p-value $< 1.23e$-2) using the proposed system compared to B2. The analysis with the objective measures reveals that the proposed BLSTM based whispered to neutral speech conversion system exhibits a superior performance in predicting smooth spectral features and excitation parameters such as voicing decisions and periodicity levels. We also observe that it exhibits comparable performance with the baseline scheme for pitch prediction.

Table. 1 provides the preference scores for the different schemes from the listening test. From the table, we find that the listeners chose the proposed BLSTM based system as the best option among all methods, 61.05% of the time. We find that the preference score of B2 is higher than that of B1 which indicates the superiority of using STRAIGHT over MLSA for synthesis. We also observe that in 4.07% of the trials the listeners chose none of the methods as the best. Across subjects, we find that the absolute increase in the percentage of trials in which the proposed BLSTM based system is chosen as the best among three schemes is 77.78%, 9.26%, 27.41%, 16.67% and 36.29% for subjects F1, F2, F3, M1 and M3, respectively. Interestingly, we find that for subject M2, where the performance of the f_0-BLSTM is worse than the model in B2 (Fig. 6 (A)), speech synthesized using B2 is chosen as the best for 7.78% (absolute) times more than the proposed method. This underscores the importance of accurate pitch prediction to synthesize natural sounding neutral speech from whispered speech. Feedback from listeners revealed that they preferred the synthesized speech which sounded the least discontinous. Such a perception, we hypothesize, could be attributed to the smoothness of the predicted spectral features. We, therefore, analyze the low pass nature of the predicted MCEPs by adopting the method proposed by Ghosh *et. al* [25], where the cutoff frequency within which 90% of the trajectory's energy is preserved is found. Hence, lower is the cutoff frequency, smoother is the trajectory. We compute the cutoff frequency for each of the 26 MCEP coefficients for every test utterance in each fold for all six subjects. We find that for most of the coefficients, the trajectory predicted by the BLSTM model is statistically significantly smoother (p-value $< 4.08e$-2) than that predicted by the baseline scheme, in more than 70% of the 24 folds (6 subjects ×4 folds) with the exception of coefficients 4, 5, 6, 7, 9 and 15. Fig. 7 provides the normalized histogram of the cutoff frequencies of the $C[0]$ trajectory predicted by P-BLSTM and B2, across all folds, for each of the six subjects. From the figure, we observe that the cutoff frequency is, on average, lower when $C[0]$ is predicted by the BLSTM model. This indicates that the MCEP coefficients, including the prosodic information carrying trajectory $C[0]$ [12], is temporally better modeled by the BLSTM.

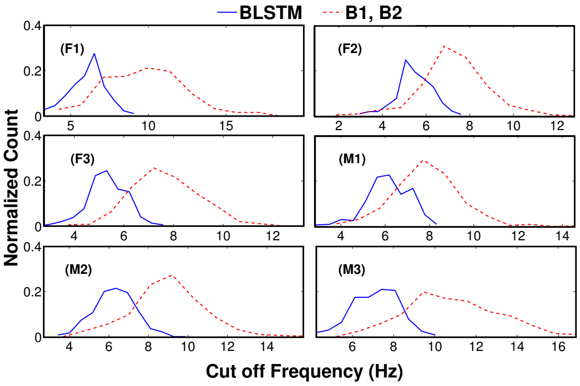


Figure 7: *Normalized histogram of the cutoff frequencies corresponding to the 0^{th} cepstral coefficient ($C[0]$) predicted using the proposed (blue) and baseline (red dashed) scheme.*

6. Conclusion

In this work, we propose a BLSTM based whisper to neutral speech conversion system, that yields temporally smoother spectral features compared to a DNN. Similarly, we observe that a BLSTM based excitation parameter prediction of voicing decisions and periodicity levels is superior to the baseline scheme considered in the study. We find that the performance of the pitch prediction using the proposed scheme is comparable to that using the baseline scheme. The analysis of the objective measures of the predicted neutral spectral features and excitation parameters from whispered speech, in conjunction with the results of a subjective listening test reveal that the proposed BLSTM based system synthesizes a more natural sounding neutral speech from whispered speech, compared to the baseline schemes. Further investigation is required to extend the proposed framework for the conversion of pathological whispered speech.

7. Acknowledgement

We thank the subjects and listeners for their participation, Ms. Kausthubha N.K for helping with the manual processing of the recordings and the Pratiksha Trust for their support.

8. 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, "What's in a whisper?" *J. Acoust. Soc. Amer.*, vol. 86, pp. 1678–1683, 1989.

[3] B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg, "Silent speech interfaces," *Speech Commun.*, vol. 52, no. 4, pp. 270–287, Apr. 2010. [Online]. Available: http://dx.doi.org/10.1016/j.specom.2009.08.002

[4] 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.

[5] I. Eklund and H. Traunmüller, "Comparative study of male and female whispered and phonated versions of the long vowels of swedish," *Phonetica*, vol. 54, no. 1, pp. 1–21, 1997.

[6] N. J. Lass, K. R. Hughes, M. D. Bowyer, L. T. Waters, and V. T. Bourne, "Speaker sex identification from voiced, whispered, and filtered isolated vowels," *The Journal of the Acoustical Society of America*, vol. 59, no. 3, pp. 675–678, 1976.

[7] 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.

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

[9] 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.

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

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

[12] 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.

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

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

[15] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3-4, pp. 187–207, 1999. [Online]. Available: https://doi.org/10.1016/S0167-6393(98)00085-5

[16] Z. Yu, V. Ramanarayanan, D. Suendermann-Oeft, X. Wang, K. Zechner, L. Chen, J. Tao, A. Ivanou, and Y. Qian, "Using bidirectional lstm recurrent neural networks to learn high-level abstractions of sequential features for automated scoring of non-native spontaneous speech," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Dec 2015, pp. 338–345.

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

[18] 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

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

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

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

[22] 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.

[23] S. Imai, "Cepstral analysis synthesis on the mel frequency scale," in *ICASSP. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 8, Apr 1983, pp. 93–96.

[24] J. Kominek, T. Schultz, and A. W. Black, "Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion," in *SLTU*. ISCA, 2008, pp. 63–68.

[25] P. K. Ghosh and S. Narayanan, "A generalized smoothness criterion for acoustic-to-articulatory inversion," *The Journal of the Acoustical Society of America*, vol. 128, no. 4, pp. 2162–2172, 2010. [Online]. Available: http://asa.scitation.org/doi/abs/10.1121/1.3455847