INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Conversion of airborne to bone-conducted speech with deep neural networks

Michael Pucher[1], *Thomas Woltron*[2]

[1]Acoustics Research Institute (ARI), Austrian Academy of Sciences (ÖAW), Vienna, Austria
[2]University of Applied Sciences Wiener Neustadt, Wiener Neustadt, Austria
michael.pucher@oeaw.ac.at, thomas.woltron@fhwn.ac.at

Abstract

It is a common experience of most speakers that the playback of one's own voice sounds strange. This can be mainly attributed to the missing bone-conducted speech signal that is not present in the playback signal. It was also shown that some phonemes have a high bone-conducted relative to air-conducted sound transmission, which means that the bone-conduction filter is phone-dependent. To achieve such a phone-dependent modeling we train different speaker dependent and speaker adaptive speech conversion systems using airborne and bone-conducted speech data from 8 speakers (5 male, 3 female), which allow for the conversion of airborne speech to bone-conducted speech. The systems are based on Long Short-Term Memory (LSTM) deep neural networks, where the speaker adaptive versions with speaker embedding can be used without bone-conduction signals from the target speaker. Additionally we also used models that apply a global filtering. The different models are then evaluated by an objective error metric and a subjective listening experiment, which show that the LSTM based models outperform the global filters.

Index Terms: voice conversion, bone-conducted speech, LSTM, deep neural networks

1. Introduction

The alienating effect of self-speech is mainly due to the missing bone-conducted signal in playback. An additional effect can be attributed to the fact that one is hearing a playback of one's own voice while not speaking, where the act of speaking as a physical and psychological process which is missing in the playback where the own voice is detached from the body. This second effect was probably much stronger with the first playback devices but might still be present in today's technical world. The bone-conducted speech is missing in most playbacks since recordings are based on airborne speech only with standard microphones, and playback of airborne and bone-conducted sound is only valid for one's own voice.

It has been shown that the alienating effect of missing bone-conducted speech leads to a short [1], but immediate negative emotional reaction [2]. There is a discrepancy of what is heard and what is expected verbalized as "too high pitched", "too nasal", "I sound so shy", or "I sound like my mother".

Furthermore listening to one's own airborne voice can lead to better learning outcomes, higher success rates, and higher endurance in blind children [3, 4], less cognitive interference and load [5, 6], is perceived as being more attractive even when not recognized as the own voice [7], and contributes significantly in perception of real and virtual environments [8]. Furthermore there are studies that investigate how children's own speech allows them to learn to distinguish between their own and other's voices [9]. [10] shows that there is a certain right-hemisphere advantage for self-compared to other-voice recognition similar to what was observed for self-face recognition.

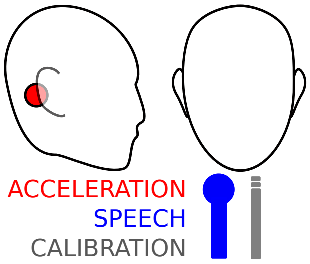

Figure 1: *Positions of the accelerometer (red), the speech recording (blue), and the calibration microphone (grey).*

Given the positive effect of listening to one's own voice and the negative effect of hearing one's own air-borne speech we think that there are several scenarios where an accurate modeling of one's own voice including bone-conducted speech is beneficial, such as Virtual Reality (VR) avatar scenarios, language learning, and music performance settings.

[11] analyzed the phonetic content of bone-conducted speech, showing that several phonemes out of ten selected phonemes have a high bone-conducted relative to air-conducted sound transmission.

Concerning the reconstruction of bone-conducted speech [12] investigates preferred equalizer settings for highest comfort when listening to one's own voice through headphones while speaking. [13] shows how to simulate one's own voice in a two-parameter model with a focus on professional singing voices. They show how to approximate one's own voice with trained singers as subjects, where bone-conduction is not singled out. [14] describes a voice conversion model for estimation of transfer characteristic in auditory feedback based on a global parameter setting. [15] describes a system for derivation of a global filter to reconstruct speech from a bone-conducted speech signal.

Due to the phone dependent bone filter a global reconstruction approach as proposed in previous work is not sufficient to model the bone-conducted speech part. Therefore we apply methods from voice conversion to convert the airborne speech signal to bone-conducted speech. These conversion methods based on LSTM recurrent deep neural networks apply a time dependent modeling where the conversion is done separately for each time frame, thereby indirectly realizing also a phone dependent model. Through the use of speaker embedding we also realize LSTM based systems where no bone-conduction signals of the target speaker are needed.

Our approach also has the potential to be used with one's own synthetic voice, where the bone-conducted part can be added to the synthetic voice via voice conversion. The usage of synthetic voices is especially interesting in VR and language

http://dx.doi.org/10.21437/Interspeech.2021-473

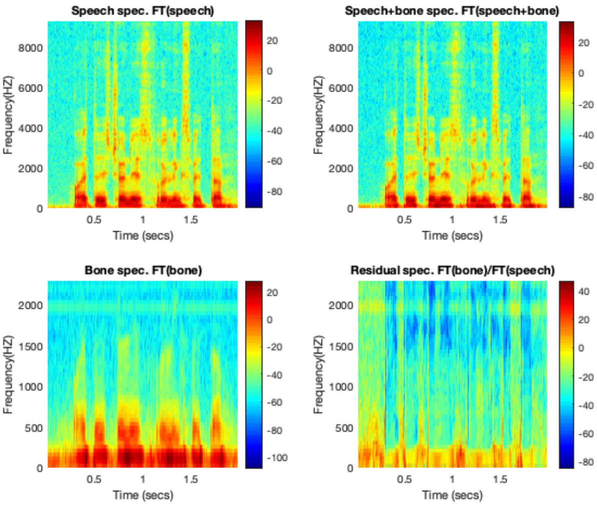

Figure 2: *Magnitude spectrogram of airborne speech (upper left), speech+bone (upper right), bone (lower left), and bone residual (lower right). FT denotes the Fourier transform. Upper figures show the full frequency range, lower figures show a restricted range that is relevant for bone-conduction.*

learning scenarios.

2. Recording and analysis

2.1. Recording

The Recording equipment for the bone-conduction included a Brüel & Kjaer accelerometer Deltatron 4507-B 005, a Metra M32 power supply and conditioning module, and a National Instruments USB data acquisition card NI-USB 6215. The acceleration signal was recorded using a Matlab-Script (R2019) under Windows 10.

For the airborne speech we also recorded a calibration track as indicated in Figure 1. The microphones for speech and calibration were plugged into a Focusrite Scarlett 18i8 Interface. The tracks were recorded with Reaper 6.02 under Windows 10.

2.2. Corpus

The recordings involved 8 non-professional speakers of Regional Standard Austrian German (RSAG). All of which gave their written consent and none reported any hearing impairment. The speakers' age ranged from 20 to 55 with 3 female speakers and 5 males. For the recordings the *Berlin* and *Marburg* corpora were used. Each corpus contains 100 sentences that are phonetically balanced with respect to the German language. They were proposed by [16] and have been used frequently since then (e.g. [17, 18]). All participants were instructed to read each sentence in a neutral voice with a small break and to repeat the sentence in case of errors, misspeaks, coughs, etc. With the two corpora and 8 speakers, a total of 1600 sentences were recorded.

2.3. Signal processing model

Figure 2 shows the magnitude spectrograms of the different signals we are dealing with. The upper two spectra are shown in a frequency range up to 9374 Hz, the lower spectra up to 2343,5 Hz since the bone-conduction related signals mainly contain energy in the lower frequency range.

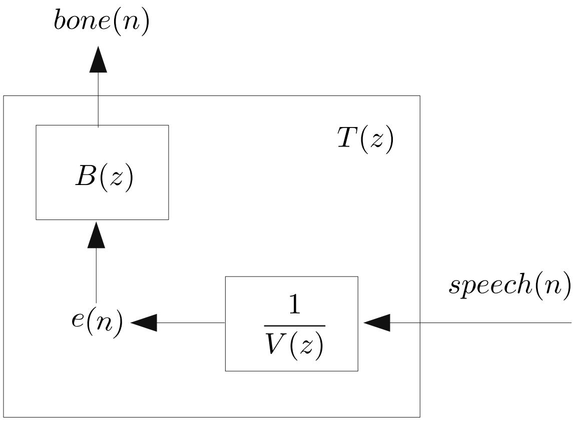

Figure 3: *System description of speech to bone conversion. The speech signal is inversely filtered to get the excitation signal, which is then filtered with the bone filter to get the bone signal.*

The system description for transforming the speech signal to the bone signal is given in Figure 3 where $T(z)$ is the overall system consisting of the inverse vocal tract filter $\frac{1}{V(z)}$ and the bone filter $B(z)$.

The speech signal is the input we can capture by a standard recording setting with a microphone. The bone-conduction signal is recorded with our bone-conduction microphone that was described previously. The mixture of speech+bone-conduction signal is the signal that approximates what a speaker hears when speaking. It is only an approximation since the bone-conduction has effects in the outer, middle, and inner ear, which cannot all be measured with our setting. Furthermore, the speech signal is recorded with a microphone in front of the speaker, thus not taking the path from mouth to ear into account. The bone residual spectrum (Figure 2, lower right) is phone dependent, therefore a simple global filtering is not sufficient for speech to bone conversion.

2.4. Spectrum analysis

We performed a decision tree based clustering, by training speaker dependent models on the bone, speech, and residual Mel-cepstra (see Figure 2, lower right) with HTS [19], using the following five questions: left-left-phone, left-phone, current-phone, right-phone, right-right-phone. For this clustering we get a lower number of used questions for bone/speech residual cepstrum as shown in Table 1. Looking at the set relations (Table 2) between bone and bone/speech trees we see a large overlap between both tree types and only a small difference in questions in bone/speech.

Although the decision tree cannot be used as a direct analysis of the phonetic content of the signal it indicates that some phone and therefore time related dynamic information is present in the residual spectra, as was already shown by [11] for a limited set of phones.

3. Conversion

3.1. Voice conversion models

Our goal is to convert an airborne speech signal into a bone-conduction signal that best matches a bone-conduction test signal that we have recorded for the target speaker. We treat this

Table 1: *Decision tree question counts (mean, minimum, and maximum value) for Mel-cepstral decision trees for the center state of a 3 state HMM trained with HTS [19] on speech, bone-conduction, and residual cepstra. Speaker dependent HMMs were trained for the 8 speakers.*

Signal type	Mean	Min	Max
Speech	75	50	88
Bone	81	67	97
Bone/speech	45	19	70

Table 2: *Mean set counts for the bone vs. bone/speech decision trees. "Diff" shows the set difference between bone/speech-bone and bone-bone/speech.*

	Intersect	Union	Diff
Bone/speech vs. bone	83.75	43.12	2.75/37.8

problem as a voice conversion problem [20]. In principle, the conversion can be performed in two ways, by either reconstructing the bone signal directly or reconstructing the bone residual spectrum and then the bone signal. In this paper we focus on the direct conversion from airborne to bone-conduction signal. For the conversion from airborne input to bone signal we used the following models.

- **Filt spk**: Speaker-dependent filter that applied one single transfer function derived from the entirety of the target speaker's airborne and bone-conduction signals to the same speaker's airborne signals.
- **Filt all**: Speaker-dependent filter that applied one single transfer function derived from the mean of all other speaker's transfer functions obtained from *Filt spk* except the target speaker's, is applied to the target speaker's airborne signals.
- **LSTM spk**: Speaker-dependent LSTM voice conversion model where data from the target speaker is used to train a model.
- **LSTM sim**: A d-vector based speaker embedding [21] is used to find the speaker in the database that is most similar to the target speaker and use an adapted model from that similar speaker for the conversion. The target speakers bone or airborne signal was not used for training the speaker independent model.
- **LSTM adapt**: Speaker-independent model is trained and adapted with data from the target speaker. The target speakers bone or airborne signal was not used for training the speaker independent model, but was used in the adaptation.
- **LSTM embed**: Speaker embedding is used in the training of a speaker-independent model, which is then used directly with airborne input and embedding from the target speaker. The target speaker's bone or airborne signal was not used for training the speaker independent model.
- **LSTM embed a.**: Speaker-independent model with speaker embedding is trained and adapted with data from the target speaker. The target speaker's bone or airborne signal was not used for training the speaker-independent model, but was used in the adaptation.

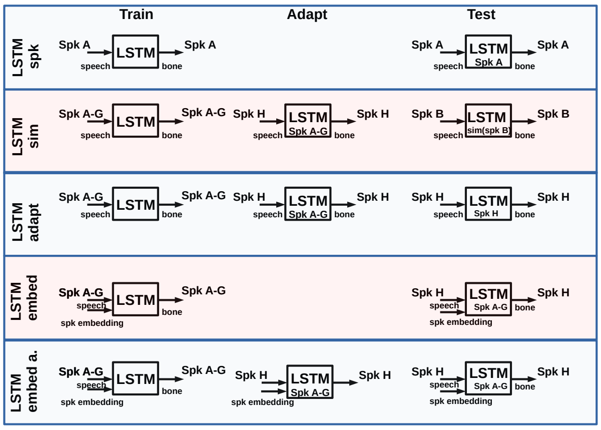

Figure 4: *LSTM based systems for conversion of speech to bone-conducted speech. Two systems (red background) do not need bone signals from the target speaker. The inputs are always features of the speech signal (plus speaker embedding), the outputs are features of the bone-conduction signal.*

The above models have different properties concerning their practical usefulness. Especially interesting are models where no bone-conduction signals of a speaker are necessary to apply the model, since these signals are difficult and costly to acquire. These are the models *Filt all*, *LSTM sim*, and *LSTM embed*. Figure 4 shows the different LSTM based conversion systems where the systems with a red background are systems where no bone-conduction signals from the target speaker are needed at test time.

3.2. Data and training

From each of the 8 speakers (5 male, 3 female) we had 200 sentences of speech and synchronous bone-conduction recordings. The data was split into 176 training, 12 development, and 12 test sentence pairs.

The features were extracted with the World vocoder [22]. 55 Mel-cepstral parameters and 5 band-aperiodicity parameters from speech were converted into bone parameters of the same type. F0 parameters were used from the source speaker.

The LSTM was a 2-layer bi-directional network [23] with 256 units, where the Pytorch implementation was used. Differently to [23] we also converted band-aperiodicty features and the 0-th Mel-cepstral component since there are differences between airborne and bone-conducted speech. Networks were trained with 150 epochs, a batch size of 12, learning rate 0.001. 100 epochs were used for adaptation. We were using a rather simple network that could be quickly trained for all models. In total we trained 64 models, 8 models for each of the two systems where no adaptation is needed, and 16 models for each of the three systems were adaptation is needed (Figure 4). In the adaptation case we need one average model for each speaker, where no data from that speaker is used.

For the speaker embedding a Residual Network (ResNet) [24] based speaker classifier was trained with 8-dimensional embedding features and 51 background speakers from our database Austrian German speakers. Embedding vectors were then generated for the 8 speakers.

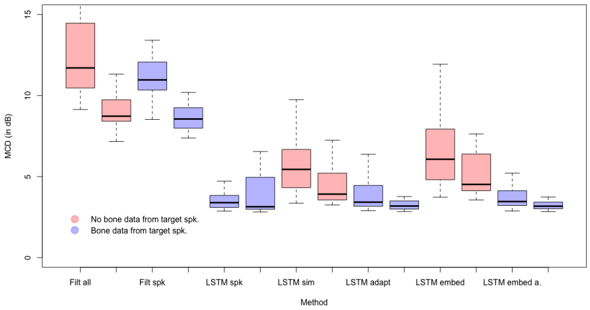

Figure 5: *Mel-cepstral distortion for the different systems. There are two bars for each method, the left bar is for all 35 Mel-cepstral coefficients, the right bar for 34 Mel-cepstral coefficients without the 0-th coefficient. Red bars are systems were no bone signals from the target speaker are needed.*

4. Evaluation

4.1. Mel-cepstral distortion

Figure 5 shows the Mel-cepstral distortion for the different systems. The two leftmost signals are the speaker-dependent and independent filters, which show a higher Mel-cepstral distortion (MCD) than all LSTM based systems. There are two bars for each system where the right bar shows the error without the 0-th Mel-cepstral coefficient. This coefficient is the average log-energy and is relevant when mixing the reconstructed bone signal with the airborne speech signal, but is less relevant for spectral speaker specific information. Therefore the left bar shows the overall error, the right bar the more speaker specific spectral error.

Looking at the overall error (left bar) a paired Wilcoxon test shows that all systems are significantly different ($p < 0.005$) except the pairs *Filt all - Filt spk*, *LSTM spk - LSTM embed a.*, *LSTM sim - LSTM embed*, *LSTM adapt - LSTM embed a.*. This shows that the trained LSTM based systems clearly outperform the simple filters and that among the systems where no bone signal from the target speaker is needed *LSTM sim* and *LSTM embed*, the systems using speaker embedding for finding a similar speaker or in training, are equally good and better than the filter-based system *Filt all*. If we exclude the 0-th coefficient the significances stay the same except that also *LSTM spk - LSTM adapt* are not significantly different.

4.2. Listening test

In a subjective listening experiment we asked 11 listeners to judge which signal out of two signals is more similar (overall) to the recorded bone-conduction signal. In this experiment we only used the three methods that are usable without bone-conduction training or adaptation data from the target speaker, which are *LSTM sim*, *LSTM embed*, and *Filt all*.

A Wilcoxon test shows that both *LSTM sim* and *LSTM embed* are significantly different ($p < 0.005$) from *Filt all* concerning this evaluation as can be also seen in Figure 6 (“Wins per listener” shows how often a method was chosen as closer to the original per listener). *LSTM sim* and *LSTM embed* are not significantly different. This shows that the two LSTM based systems outperform the filter-based system in the subjective ratings, which is very promising for using these methods in future applications.

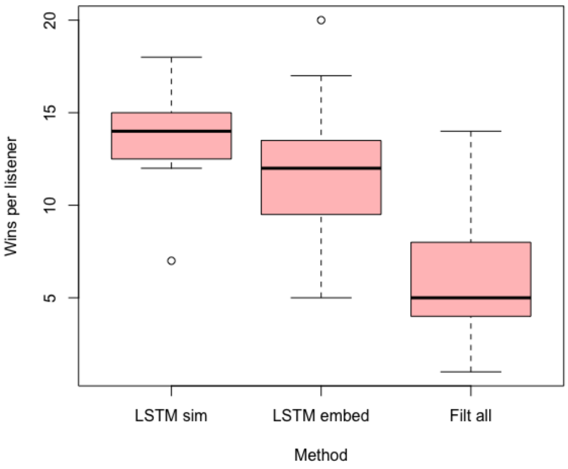

Figure 6: *Results per method from a listening test with 11 listeners. The y-axis shows the number of wins per method for the listeners, the x-axis shows the three methods that can be used without any bone-conduction data from the target speaker. The methods have been compared with the recorded bone-conducted speech signal.*

5. Conclusion

We presented different systems for the conversion from airborne to bone-conducted speech and showed in an objective evaluation that all LSTM based systems outperform the filter-based systems. This means that the LSTM based systems are better in capturing the dynamic spectral part that is present in the bone-conduction signal but not in the airborne signal.

In a subjective evaluation with the three systems that can be trained without bone-conducted speech from the target speaker we also saw that the LSTM based systems are judged as producing signals that are more similar to the recorded bone-conducted speech. These objective and subjective evaluation results are very promising for applications of this conversion in singing education, language learning, and VR.

Concerning the evaluation there is still a need to evaluate how the bone-conducted speech and the speech with added reconstructed bone-conduction is perceived by the recorded speakers themselves when listening to their own speech. These evaluations need to be done in the lab with additional equipment to simulate the bone-conduction hearing in the listeners. Listeners will be asked to judge different signals or we will measure relevant bio-signals of the listeners or task completion times.

The purpose of the current study was to develop a method for reconstructing bone-conducted speech as it can otherwise be measured from outside the speaker’s head with accelerometers. We are fully aware that this cannot entirely reflect the speaker’s auditory impression with the actual bone-conducted speech’s multiple pathways reaching the inner ear together with other sensory impressions a speaker might have. Despite the limitations, the results of our approach show that it is possible to reconstruct the bone-conducted component of speech without the need to record it individually for each speaker.

6. References

[1] P. Holzman, C. Rousey, and C. Snyder, "On listening to one's own voice: effects on psychophysiological responses and free associations," *Journal of Personality and Social Psychology*, vol. 4, no. 2, pp. 432–441, 1966.

[2] M. Daryadar and M. Raghibi, "The effect of listening to recordings of one's voice on attentional bias and auditory verbal learning," *International Journal of Psychological Studies*, vol. 7, 06 2015.

[3] M. Pucher, M. Toman, D. Schabus, C. Valentini-Botinhao, J. Yamagishi, B. Zillinger, and E. Schmid, "Influence of speaker familiarity on blind and visually impaired children's perception of synthetic voices in audio games," in *Proc. Interspeech*, Dresden, Germany, 2015, pp. 1625–1629.

[4] M. Pucher, B. Zillinger, M. Toman, D. Schabus, C. Valentini-Botinhao, J. Yamagishi, E. Schmid, and T. Woltron, "Influence of speaker familiarity on blind and visually impaired childrens and young adults perception of synthetic voices," *Comput. Speech Lang.*, vol. 46, no. C, p. 179–195, Nov. 2017.

[5] P. Tacikowski, T. Freiburghaus, and H. H. Ehrsson, "Goal-directed processing of self-relevant information is associated with less cognitive interference than the processing of information about other people." *Journal of Experimental Social Psychology*, vol. 68, p. 93–100, 2017.

[6] J. Graux, M. Gomot, and S. Roux, "My voice or yours? an electrophysiological study," *Brain Topog*, vol. 26, p. 72–82, 2013.

[7] S. Hughes and M. Harrison, "I like my voice better: Self-enhancement bias in perceptions of voice attractiveness," *Perception*, vol. 42, pp. 941–9, 09 2013.

[8] C. Pörschmann, "Influences of bone conduction and air conduction on the sound of one's own voice," *Acta Acustica united with Acustica*, vol. 86, no. 6, pp. 1038–1045, 2000.

[9] C. Fernyhough and J. Russell, "Distinguishing one's own voice from those of others: A function for private speech?" *International Journal of Behavioral Development*, vol. 20, no. 4, pp. 651–665, 1997.

[10] C. Rosa, M. Lassonde, C. Pinard, J. P. Keenan, and P. Belin, "Investigations of hemispheric specialization of self-voice recognition," *Brain and Cognition*, vol. 68, no. 2, pp. 204–214, 2008.

[11] S. Reinfeldt, P. Ostli, B. Håkansson, and S. Stenfelt, "Hearing one's own voice during phoneme vocalization—transmission by air and bone conduction," *The Journal of the Acoustical Society of America*, vol. 128, pp. 751–62, 08 2010.

[12] O. Lindholm, "Preferred EQ-setting for highest comfort when listening to one's own voice through headphones while speaking," Master's thesis, Luleå University of Technologys, 2014.

[13] S. Y. Won, J. Berger, and M. Slaney, "Simulation of one ' s own voice in a two-parameter model," in *Proc. of the Int. Conf. Music Perception and Cognition*, 2014.

[14] S. Morita, D. Kawamoto, and T. Toya, "Voice conversion model for estimation of transfer characteristic in auditory feedback," in *ICA*, 2019, pp. 6630–6636.

[15] T. Tamiya and T. Shimamura, "Reconstruction filter design for bone-conducted speech." in *Proc. Interspeech*, Jeju Island, Korea, 2004, pp. 1085–1088.

[16] J. Sotschek, "Sätze für sprachgütemessungen und ihre phonologische anpassung an die deutsche sprache," *Tagungsband DAGA: Fortschritte der Akustik*, pp. 873–876, 1984.

[17] K. Fellbaum, "Sprachqualitätsmessungen," in *Sprachverarbeitung und Sprachübertragung*. Springer, 2012, pp. 127–172.

[18] M. Pätzold and A. P. Simpson, "Acoustic analysis of German vowels in the kiel corpus of read speech," *Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung Universität Kiel*, vol. 32, pp. 215–247, 1997.

[19] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0." in *SSW*. Citeseer, 2007, pp. 294–299.

[20] B. Sisman, J. Yamagishi, S. King, and H. Li, "An overview of voice conversion and its challenges: From statistical modeling to deep learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 132–157, 2021.

[21] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 4052–4056.

[22] M. Morise, F. Yokomori, and K. Ozawa, "World: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. E99.D, pp. 1877–1884, 07 2016.

[23] L. Sun, S. Kang, K. Li, and H. Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4869–4873.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.