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

Pitch contour separation from overlapping speech

Hiroki Mori

Utsunomiya University, Japan
hiroki@speech-lab.org

Abstract

In everyday conversation, speakers' utterances often overlap. For conversation corpora that are recorded in diverse environments, results of pitch extraction in the overlapping parts may be incorrect. The goal of this study is to establish the technique of separating each speaker's pitch contour from an overlapping speech in conversation. The proposed method estimates statistically most plausible f_o contour from the spectrogram of overlapping speech, along with the information of the speaker to extract. Visual inspection of the separation results showed that the proposed model was able to extract accurate f_o contours from overlapping speeches of specified speakers. By applying this method, voicing decision errors and gross pitch errors were reduced by 63 % compared to simple pitch extraction for overlapping speech.

Index Terms: prosody, conversation, source separation, neural f_o model

1. Introduction

Prosody, which refers to the unique aspects of spoken language such as intonation, rhythm, voice quality, etc. is a central theme of the speech science. Among these aspects, pitch is most important. It is widely known that pitch is related to almost all aspects of speech communication, e.g. lexical accent and intonational phonology [1], languages and dialects [2], speaker individuality including age and gender [3], paralinguistics [4], social signals [5], and spoken language processing.

Today, corpus-based linguistics has become the norm. However, it took the advent of a large-scale spontaneous speech corpus, such as Corpus of Spoken Japanese (CSJ), to make corpus-based phonetics realistic. The CSJ, and other corpora as well, provide a rich collection of speech that is suitable to investigate a wide variety of speech phenomena. Nevertheless, they are not definitely suitable for studying speech in conversation. The main reason for this is that conversations in such corpora were "set up" exclusively for the corpus construction. On the other hand, collecting everyday conversations, rather than set-up ones, is ideal from the perspective of ecological validity. The project of building the *Corpus of Everyday Japanese Conversation (CEJC)* [6], is a major challenge that will bring breakthroughs in the studies of speech and conversation. The CEJC focuses on conversations embedded in naturally occurring activities in daily life, without the exogenous intervention of researchers imposing topics or displacing the context of action [7]. It contains more than 200 hours of conversations of various types that enable to capture the diversity of everyday conversations and to observe natural conversational behavior in our daily life.

In its individual-based recording, informants carried portable recording devices and recorded her/his everyday activities in a variety of locations such as at home, office, or restaurants, etc. Because of the diverse environments, the quality of the recording is not always satisfactory. The main problem in analyzing the prosody of speech in the corpus is that each speaker's voice is not acoustically separated. In everyday conversation, speakers' utterances often overlap, and the results of pitch extraction in the overlapping parts may be incorrect. The inaccurately estimated pitch contour makes it impossible to analyze the prosody of conversational speech.

The goal of this study is to establish the technique of separating each speaker's pitch contour from an overlapping speech in conversation. This technique allows us to quantitatively analyze the prosody of everyday conversation, which current technology cannot handle. The method applies the idea of f_o contour generation in neural speech synthesis, where the neural f_o model learns natural f_o patterns that human can generate from a large amount of speech data, and predicts statistically most plausible f_o contour under the constraints of the given linguistic information. The proposed pitch contour separation method provides the model with the acoustic signal of overlapping speech instead of linguistic information, along with the information of the speaker to extract.

2. Related studies

The pitch contour separation is considered as a part of a broader class of problems called multi-pitch analysis. Its major application is the automatic transcription of music [8]. Various approaches have been proposed, including iterative estimation on the frequency domain [8], statistical modeling of multiple harmonic structures [9], nonnegative matrix factorization with harmonic constraints [10], and the "specmurt" analysis [11]. Another aspect of the multi-pitch analysis is the computational auditory scene analysis for simultaneous speech [12]. A neural comb filtering [13] is a typical work that has attempted to estimate both pitches of concurrent natural speech.

The multi-pitch analysis is closely related to sound source separation. Source separation techniques can be classified into two categories: single-channel and multi-channel. In this paper, the input to the model is assumed to be single-channel. However, in most recordings in the CEJC, each participant wore her or his own IC recorder, so the recording was multi-channel. As described later, the proposed method can be easily extended to multi-channel input.

As f_o models for text-to-speech synthesis, recurrent neural networks (RNNs) have been successfully used [14]. To make better use of temporal dependency in f_o contours, autoregressive neural f_o models were proposed [15, 16]. In [16], discretizing f_o values was proposed, as in the speech waveform modeling by WaveNet [17], and claimed that the discrete f_o models generated accurate and less oversmoothed f_o contours compared to the continuous f_o models.

http://dx.doi.org/10.21437/Interspeech.2021-2164

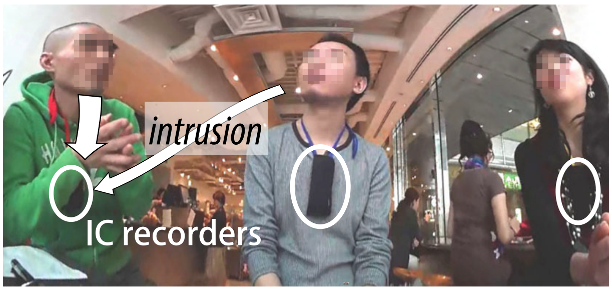

Figure 1: *Typical recording situation of the CEJC.*

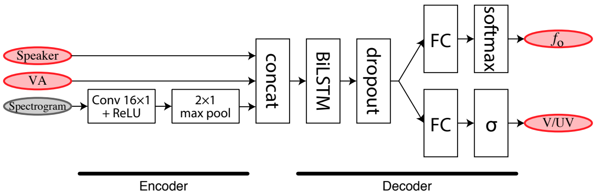

Figure 2: *The architecture of proposed model.*

3. The goal

In the typical recording situations of the CEJC, each participant wore an IC recorder, in addition to the one placed in the center. By using individual recording devices, the level of each speaker's voice becomes dominant in each recording device. However, this cannot completely prevent other speakers' voices from intruding into the recording, resulting in an overlapping speech (Fig. 1). The goal of this study is to separate each speaker's f_o contour from the overlapping speech.

The current study assumes:

- Utterances of all the speakers are recorded by a central IC recorder without being separated. To make the problem simpler and more general, multi-channel recordings are not used in the current study.
- Identities of present speakers are known for given overlapping speech.
- The presence/absence of each speaker's utterances, in other words, speech activities, is known. The CEJC provides the information of manually-annotated utterance unit including startTime / endTime, so this assumption is reasonable for analyzing the prosody of the CEJC.
- The objective is to estimate the f_o contour of each speaker's utterance from the overlapped speech.

To evaluate the precision of an estimated pitch contour, ground-truth f_o information is needed. However, this is not available for the CEJC because it does not provide clean and separated speech waveforms. Therefore, simulated overlapping speech is used in this paper. By mixing two speech waveforms of different speakers, the recording of the central IC recorder is simulated. The ground truth is assumed to be the f_o contour estimated for the original waveform before the mixing by an ordinary pitch extractor.

4. Network architecture

The architecture of proposed model is shown in Fig. 2. The input of the model is the spectrogram of an overlapping speech, represented by a $N \times D$ matrix, where N is the number of time frames and D is the number of frequency bins. A stack of convolutional layers, called Encoder, converts the spectral pattern of each frame into some pitch features, which presumably correspond to possible f_o candidates, then passes them as a vector. The vector is then combined with the target speaker's voice activity information (1/0), as well as her / his speaker embedding. For each time frame, the Decoder section, composed of an LSTM layer, a fully-connected layer, and a softmax layer, outputs a discrete f_o value.

It is easy to extend the input to the encoder for further improvement. For example, appending the waveform from the per-speaker IC recorder to the input will help the separation, because the target speaker's utterances were already enhanced.

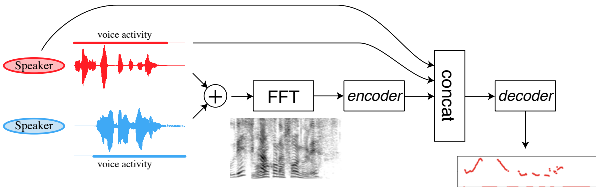

Figure 3: *Loss calculation for Speaker A.*

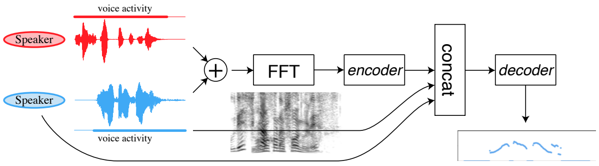

Figure 4: *Loss calculation for Speaker B.*

5. Network training

To simulate the overlapping speech, two utterances of different speakers are randomly sampled from a training corpus of clean and separate speech. The second utterance is then randomly shifted backward or forward. Finally, an overlapping speech is created by adding these waveforms. The random pairing and shifting are repeated for every epoch of training.

In the network training, the losses are accumulated for all the present speakers in the overlapping speech. The detail for a two-speaker case is described below. First of all, the spectrogram of the overlapping speech was calculated then encoded. Then, the encoded vector for each frame is concatenated with Speaker A's embedding and voice activity information. The decoder estimates the f_o contour from the encoded sequence (Fig. 3), then calculate the loss using Speaker A's ground truth. This process is repeated for the Speaker B, using the identical encoded sequence of the spectrogram (Fig. 4).

Finally, the total loss is calculated by summed up the loss for all speakers:

$$\text{Loss} = \sum_{s \in \{A,B\}} (\text{Loss}_s^{f_o} + \text{Loss}_s^{\text{VUV}}), \tag{1}$$

$$\text{Loss}_s^{f_o} = - \sum_{n \in \text{VA}_s} t_{sn}^{\text{VUV}} \log(\mathbf{y}_{sn}^{f_o})_{t_{sn}^{f_o}}, \tag{2}$$

$$\text{Loss}_s^{\text{VUV}} = - \sum_{n \in \text{VA}_s} (t_{sn}^{\text{VUV}} \log y_{sn}^{\text{VUV}} + (1 - t_{sn}^{\text{VUV}}) \log(1 - y_{sn}^{\text{VUV}})), \tag{3}$$

where $(\mathbf{y})_t$ denotes the t's element of the softmax output vector $\mathbf{y}$; $\mathbf{y}^{f_\mathrm{o}}_{sn}$ and y^{VUV}_{sn} denote f_o and voiced / unvoiced output for the speaker s at the time frame n; $t^{f_\mathrm{o}}_{sn}$ denotes the index of the discretized ground-truth f_o; t^{VUV}_{sn} denotes the ground-truth voicedness (voiced: 1, unvoiced: 0); and VA$_s$ is a set of time frame indices where speaker s's utterance is present, which is known (see Sect. 3). Note that the f_o loss in Eq. (2) is taken into account only for voiced frames.

The per-speaker decoding and loss calculation framework described above is an elegant solution that is free from the permutation problem [18], and can be easily extended to three or more speakers.

6. Evaluation

6.1. Experiment conditions

As a dataset of clean, separate speech, the Simulated Public Speaking Subset of the CSJ was adopted, from which utterances that contain disfluency such as filler or hesitation were excluded. The subset was then divided into training set consisting of 98967 inter-pausal units (IPUs) of 344 speakers, "finetune" set consisting of 7555 IPUs of 20 speakers not including the speakers for training, and test set consisting of another 3238 IPUs of the same 20 speakers as the finetune set.

The waveforms were sampled at 16 kHz. The ground-truth f_o contours were estimated from the waveforms with the YangSaf [19] algorithm with a frame shift of 5 ms. As in [16], the f_o values were quantized into 255 levels between 80 Hz and 600 Hz, equally spaced on the log scale. Its theoretical root mean squared quantization error is 0.08 semitone.

For the spectral analysis, the waveforms were windowed by a 640-point hann window with a frame shift of 5 ms, then FFTed to convert to spectra. The spectrograms were the time series of logarithmic amplitude spectra with a linear frequency scale.

The encoder was composed of a 32-channel 16 × 1 convolution layer applied with weight normalization [20], an ReLU activation, and a 2 × 1 max pooling layer. The 16 × 1 convolution was meant for capturing harmonic structures, or periodicity in a log spectrum, assuming time independence. The decoder was composed of a bi-directional LSTM layer with a dimension of 512 × 2, followed by a dropout layer, then fully-connected layers for 255-level quantized f_o value and voiced / unvoiced decision.

The speaker embedding has 16 dimensions, each of which was randomly initialized. Figure 5 visualizes the distribution of learned speaker embeddings mapped to the first and second principal components. As can be seen from the clear separation of female and male distributions, the learned embeddings reflect the speaker individuality that helps to separate the individual f_o contour from the overlapping spectrogram. From the author's observation, the first and second components seemed to correspond roughly to the overall pitch height and loudness of the speaker, respectively.

In the following evaluation, the test was performed for speakers that are not included in the training set. However, the speaker embeddings for the target speakers must be known to apply the proposed method. Since the purpose of this experiment was to check the feasibility of the overall design of this method, the embeddings for target speakers were also *learned* this time. The estimation of embeddings for target speakers was performed in the following manner. First, a temporary model was trained with an extended dataset that combines the training set and the "finetune" set. Once converged, the speaker embeddings were frozen. Then, a new model was trained with the training set using the speaker embeddings. Therefore, the utterances of the target speakers were not used in the model training itself. Currently, the "finetune" set was used only for this, and no speaker adaptation was applied.

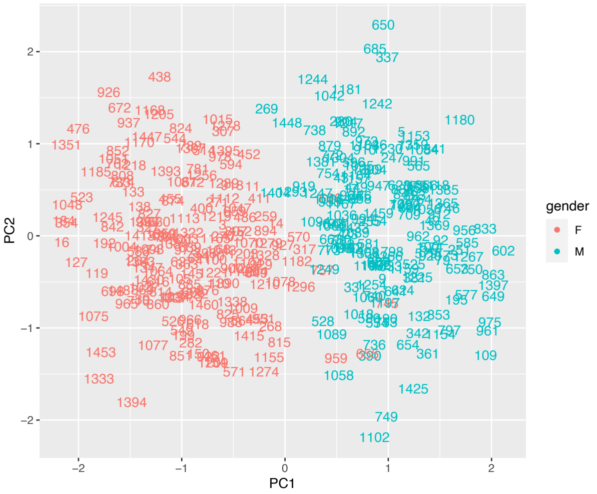

Figure 5: *Visualization of learned speaker embeddings. Red: female, Blue: male. Digits represent the speaker ID in the CSJ.*

To apply this method to another corpus, we will need to estimate the embedding of each speaker in the corpus, presumably from her / his non-overlapping speech parts. The author assumes that this is not as difficult as estimating embeddings for unseen speakers for text-to-speech systems [21, 22], because the speaker embeddings for the f_o separation are mainly related to the speaking fundamental frequency, as shown in Fig. 5. Replacing the above speaker embeddings with the ones estimated by an off-the-shelf speaker model for speaker recognition could be an option, but its effectiveness needs to be examined separately.

6.2. Results

The proposed method worked great for most cases. Figures 6 to 8 illustrate several examples, each of which shows the ground-truth f_o contour before mixing, the estimated f_o contour for one speaker, and the one for another speaker. In the case shown in Fig. 6, the final rising boundary tone of the female speaker is successfully preserved, even though the part is overlapped by the beginning of the male speaker. Figure 7 shows another example of female and male speakers with interwoven f_o contours. In this example, the ground-truth f_o of the male speaker is wrong (pitch doubling) in the vicinity of 1.1 to 1.2 sec, due to his creaky voice quality at the utterance final. The proposed model even corrected such pitch extraction errors due to its powerful f_o modeling capabilities. Finally, Fig. 8 shows the result for a case of two male speakers. Although the two f_o contours are very close to each other, the proposed method separated them quite well.

The accuracy of the proposed pitch contour separation method was assessed using standard measures for the evaluating pitch extraction algorithms [23, 24], as described below:

Voicing Decision Error (VDE) The proportion of frames for which the voiced / unvoiced decision is wrong.

Gross Pitch Error (GPE) The proportion of frames where the

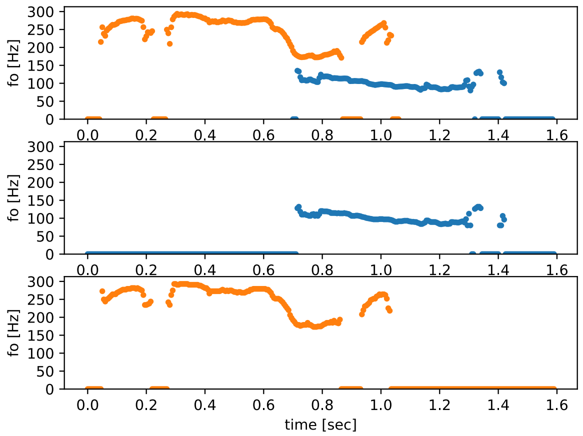

Figure 6: *True and estimated f_o contours (1).*

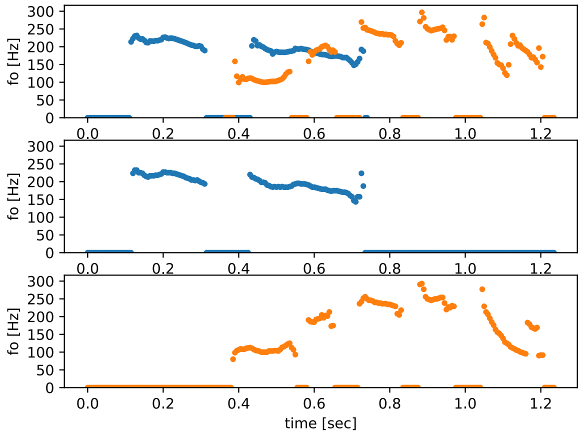

Figure 7: *True and estimated f_o contours (2).*

estimated pitch is drastically different from the ground truth. In this paper, an error was considered to be gross when the difference between the estimated and ground-truth f_o exceeds 10 % ($= \pm 1.65$ semitones).

Fine Pitch Error (FPE) The standard deviation of the difference between the estimated and ground-truth f_o, for which the difference does not exceed 10 %.

The errors were calculated only for the interval where each speaker's utterance is present.

Table 1 shows the result. Here, "no separation" is the result of simply applying YangSaf to the overlapping speech. For non-overlapping parts, these two conditions are not much different. For overlapping parts, however, the pitch determination for at least one speaker will always fail if not separated. With the proposed pitch separation method, the voicing decision error and gross pitch error was drastically reduced, as shown in Tab.1.

7. Conclusions

In this paper, a novel pitch contour separation method from overlapping speeches was introduced. A neural network consisting of a convolutional network and an RNN was trained to estimate the pitch contour of a specified speaker from the spectrograms of artificially mixed utterances of two speakers. The network learns what the pitch pattern of human speech looks like and, using the specified speaker embedding as a clue, finds the most probable f_o among the candidates distilled from the spectrum of mixed speech. Visual inspection of the separated pitch contours suggests that the proposed method is very promising for prosodic analysis of overlapping speech. A quantitative evaluation showed that the proposed method can reduce the voicing decision errors and gross pitch errors by 63%.

Applying the proposed method to the real conversation corpus CEJC is clearly the next step. Evaluating this method for CEJC is not straightforward because the ground-truth f_o is not known — that is the very reason why the f_o contour separation is needed! To make the ground truths for evaluation, manual annotation of the pitch contours of individual speakers for the overlapping utterances will be needed. The proposed framework is also expected to assist such annotation works, by integrating with popular tools such as Praat to provide an intelligent multi-pitch tracking interface with manual correction facilities, similar to waypoints to alter routes in Google Maps.

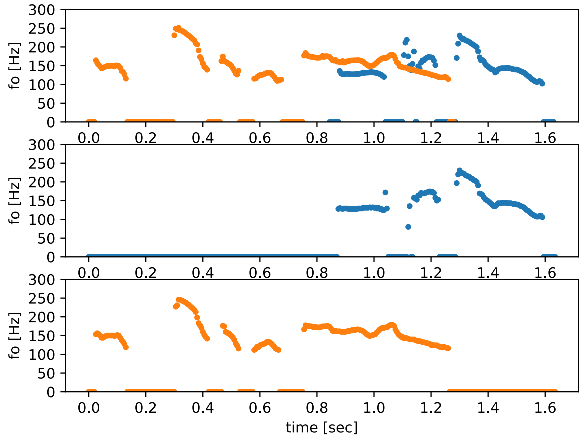

Figure 8: *True and estimated f_o contours (3).*

Table 1: *Quantitative evaluation of the proposed pitch separation method.*

	VDE (%)	GPE (%)	FPE (st)
no separation	10.72	17.50	0.33
proposed	3.99	6.56	0.34

8. Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 19H01252.

9. References

[1] D. R. Ladd, *Intonational Phonology*, 2nd ed. Cambridge University Press, 2008.

[2] M. Dolson, "The pitch of speech as a function of linguistic community," *Music Perception*, vol. 11, no. 3, pp. 321–331, 1994.

[3] H. Kasuya and H. Yoshida, "Age-related changes in the human voice," in *Aging Voice*, K. Makiyama and S. Hirano, Eds. Singapore: Springer, 2017, pp. 27–36.

[4] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. S. Narayanan, "Paralinguistics in speech and language: State-of-the-art and the challenge," *Computer Speech and Language*, vol. 27, no. 1, pp. 4–39, 2013.

[5] A. S. Pentland, *Honest Signals: How They Shape Our World*. London: The MIT Press, 2008.

[6] H. Koiso, Y. Den, Y. Iseki, W. Kashino, Y. Kawabata, K. Nishikawa, Y. Tanaka, and Y. Usuda, "Construction of the corpus of everyday Japanese conversation: An interim report," in *Proc. LREC 2018*, 2018, pp. 4259–4264.

[7] L. Mondada, "The conversation analytic approach to data collection," in *The Handbook of Conversation Analysis*, J. Sidnell and T. Stivers, Eds. Wiley-Blackwell, 2012, ch. 3, p. 32–56.

[8] A. Klapuri, "Signal processing methods for the automatic transcription of music," Ph.D. dissertation, Tampere University of Technology, 2004.

[9] H. Kameoka, T. Nishimoto, and S. Sagayama, "A multipitch analyzer based on harmonic temporal structured clustering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 982–994, 2007.

[10] S. A. Raczyński, N. Ono, and S. Sagayama, "Multipitch analysis with harmonic nonnegative matrix approximation," in *Proc. ISMIR 2007*, 2007, pp. 381–386.

[11] S. Saito, H. Kameoka, K. Takahashi, T. Nishimoto, and S. Sagayama, "Specmurt analysis of polyphonic music signals," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 16, no. 3, pp. 639–650, 2008.

[12] H. G. Okuno, T. Nakatani, and T. Kawabata, "Listening to two simultaneous speeches," *Speech Communication*, vol. 27, no. 3, pp. 299–310, 1999.

[13] A. de Cheveigné, "Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing," *Journal of the Acoustical Society of America*, vol. 93, no. 6, pp. 3271–3290, 1993.

[14] R. Fernandez, A. Rendel, B. Ramabhadran, and R. Hoory, "Prosody contour prediction with long short-term memory, bi-directional, deep recurrent neural networks," in *Proc. Interspeech 2014*, 2014, pp. 2268–2272.

[15] X. Wang, S. Takaki, and J. Yamagishi, "An autoregressive recurrent mixture density network for parametric speech synthesis," in *Proc. ICASSP 2017*, 2017, pp. 4895–4899.

[16] ——, "Autoregressive neural F0 model for statistical parametric speech synthesis," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 26, no. 8, pp. 1406–1419, 2018.

[17] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," *CoRR*, vol. abs/1609.03499, 2016.

[18] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.

[19] H. Kawahara, Y. Agiomyrgiannakis, and H. Zen, "Using instantaneous frequency and aperiodicity detection to estimate F0 for high-quality speech synthesis," in *Proc. ISCA Speech Synthesis Workshop (SSW9)*, 2016, pp. 238–245.

[20] T. Salimans and D. P. Kingma, "Weight normalization: A simple reparameterization to accelerate training of deep neural networks," in *Proc. NIPS '16*, 2016, p. 901–909.

[21] S. Arik, J. Chen, K. Peng, W. Ping, and Y. Zhou, "Neural voice cloning with a few samples," in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31, 2018.

[22] E. Cooper, C. I. Lai, Y. Yasuda, F. Fang, X. Wang, N. Chen, and J. Yamagishi, "Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings," in *Proc. ICASSP 2020*, 2020, pp. 6184–6188.

[23] W. J. Hess, "Pitch and voicing determination of speech with an extension toward music signals," in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. A. Huang, Eds. Berlin, Heidelberg: Springer, 2008, ch. 10.

[24] T. Drugman and A. Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics," in *Proc. Interspeech 2011*, 2011, pp. 1973–1976.