Time Alignment using Lip Images for Frame-based Electrolaryngeal Voice Conversion

Yi-Syuan Liou*, Wen-Chin Huang*†, Ming-Chi Yen*, Shu-Wei Tsai‡,
Yu-Huai Peng*, Tomoki Toda†, Yu Tsao*, Hsin-Min Wang*
* Academia Sinica, Taiwan
† Nagoya University, Japan
‡ National Cheng Kung University Hospital, Taiwan
E-mail: wen.chinhuang@g.sp.m.is.nagoya-u.ac.jp

***Abstract*—Voice conversion (VC) is an effective approach to electrolaryngeal (EL) speech enhancement, a task that aims to improve the quality of the artificial voice from an electrolarynx device. In frame-based VC methods, time alignment needs to be performed prior to model training, and the dynamic time warping (DTW) algorithm is widely adopted to compute the best time alignment between each utterance pair. The validity is based on the assumption that the same phonemes of the speakers have similar features and can be mapped by measuring a pre-defined distance between speech frames of the source and the target. However, the special characteristics of the EL speech can break the assumption, resulting in a sub-optimal DTW alignment. In this work, we propose to use lip images for time alignment, as we assume that the lip movements of laryngectomee remain normal compared to healthy people. We investigate two naive lip representations and distance metrics, and experimental results demonstrate that the proposed method can significantly outperform the audio-only alignment in terms of objective and subjective evaluations.**

I. Introduction

Laryngectomy is a common type of speech disorder, which refers to the surgery that removes the larynx including the vocal folds, as a therapy of laryngeal cancer. Patients undergone such a surgery are called laryngectomees, who lose the ability to produce source excitation and are no longer to produce speech. They often resort to a speaking-aid device called the electrolarynx (EL), which generates excitation signals outside the patient's body. These excitation signals are conducted as alternative excitation sounds into the oral cavity and articulated to produce EL speech sounds. The produced speech, which we refer to as electrolaryngeal speech (EL speech), suffers from the mechanical excitation signals and ends up robotic and unnatural compared with natural speech.

To improve the quality of EL speech, the major trend is to apply statistical voice conversion (VC) [1]–[4], a technique that converts one type of speech to another without changing the underlying contents, which we will hereafter refer to as ELVC. Typically, such a VC system consists of three stages: analysis, conversion, and synthesis. First, acoustic features are extracted from the source EL speech. Then, a statistical model trained with a parallel dataset consisting of pairs of EL speech and natural speech takes as input the source acoustic features and generates the converted acoustic features. Finally, a waveform synthesis module restores the final waveform signal. The conversion model has evolved over time, from traditional Gaussian mixture models (GMMs) [1] to recent deep neural network (DNN) models [4]–[6].

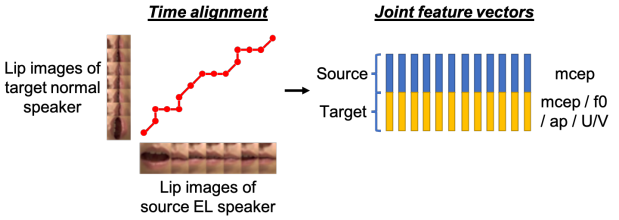

Fig. 1: Illustration of using lip images to construct joint feature vectors for frame-based electrolaryngeal voice conversion model training.

A crucial step in ELVC is the time alignment between the source EL speech and the target natural speech. In the conventional VC literature, a temporal alignment method must be employed during the training of *frame-based* models like GMM, since the joint probability density function (p.d.f.) between the source and target acoustic feature frames are modeled in a frame-synchronized manner [7]. The most widely adopted approach is the dynamic time warping (DTW) algorithm [8], which finds the optimal alignment path of two feature sequences by considering some predefined similarity measure. While a correspondence between the phonetic similarity and a simple measurement like the L2-distance between the acoustic frames is assumed in normal VC, it is however not always true in ELVC. Since the acoustic characteristics of the artificial EL speech and the natural speech are different, the similarity calculation may be inaccurate. Such sub-optimal alignments may bound the conversion performance.

There have been attempts to tackle this issue. While the DTW algorithm operates on the utterance level, [9] utilized the phonetic labels and performed DTW on the phoneme level, where the annotating process can be laborious. The labeling process can be replaced with forced alignment as in [10], but an accurate ASR model for EL speech would then be needed. Another direction is to integrate the alignment process and the model training. For example, an early attempt used a so-called DP-GMM model [11] whose convergence speed and performance suffers from the one-to-one alignment

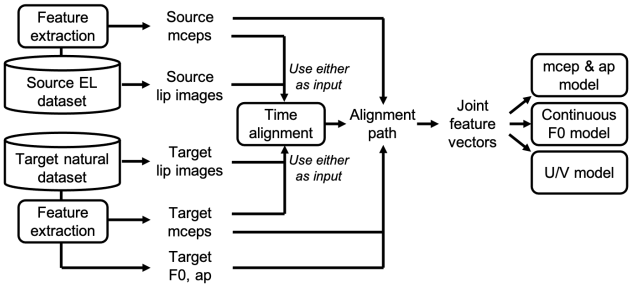

Fig. 2: Training in electrolaryngeal voice conversion.

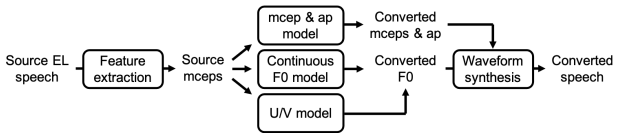

Fig. 3: Conversion in electrolaryngeal voice conversion.

assumption. Recently, a sequence-to-sequence approach was shown to be promising [12], while the complex computation limits the ability to realize real-time applications [6].

In this work, we propose to improve the accuracy of the temporal alignment procedure by leveraging the accompanied lip images when the EL speech are produced. The motivation is based on the observation that the lip movements of laryngectomees still remain normal. Despite the problem of homophones [13], where auditorily distinct sound units share almost identical lip shapes, we hypothesize that the similarity between lip images of a EL speaker and those of a natural speaker can better reflect the underlying phonetic correspondence. As shown in Figure 1, since the lip images and the speech are time-synchronized, the DTW path obtained using the lip images can be used to align the acoustic features of the EL and normal speech to train a VC model. Thus, the lip images are not required during the conversion phase. We evaluate our proposed method on an internal dataset for ELVC, and experimental results show that several aspects can be improved, including objective distortion measures and subjective quality.

II. Basic framework of electrolaryngeal voice conversion based on frame-level modeling

A. Training and conversion processes

The training process is depicted in Figure 2. To train a statistical ELVC model, assume we have access to a *parallel* training set containing pairs of normal and EL speech utterances that are of the same contents. A high quality parametric vocoder, such as WORLD [14], is first used to decompose the waveform signals into several acoustic features from the normal and EL sentences, including spectral features (specifically mel-cepstrum coefficients (mceps)), fundamental frequency (F0) and aperiodicity signal (ap). The mceps are used to perform time alignment, which then constructs the joint feature vectors. Due to the special characteristics of EL speech, only the mceps are considered normal. Therefore, following [4], conditioned on the EL spectral features, three models are separately trained to predict the segmental features (mceps and ap), continuous F0 and unvoiced/voiced symbol (U/V).

The conversion process is depicted in Figure 3. As described in Section I, three stages are performed sequentially. The mcep sequence of the input EL speech is first extracted, and is used as the input of the trained conversion models to generate the converted features. A waveform synthesizer finally generates the converted waveform with the converted features.

B. DTW based on mcep features

As a baseline, we considered an iterative alignment process based on DTW with mel-cepstrum coefficients (mceps) which we will refer to as DTW-mcep. Please note that we utilized *sprocket* [15][1], an open-source toolkit implementing GMM-based VC [7]. Sprocket was designed for VC between normal speech, and careful modifications need to be made to the alignment process for EL speech as in [4]. Nonetheless, we chose sprocket for its simplicity and reproducibility.

First, silence removal and dynamic feature extension are performed. Then, the following steps are iteratively performed:

1) The DTW algorithm minimizes a distance metric between the aligned source and target feature vectors. For mcep inputs, the mel-cepstrum distortion (MCD) is often used, whose definition is as follows:
$$MCD[dB] = \frac{10}{\log 10}\sqrt{2\sum_{d=1}^{K}(mcep_d^{(s)} - mcep_d^{(t)})^2}, \quad (1)$$
where K is the dimension of the mceps and $mcep_d^{(s)}$ and $mcep_d^{(t)}$ represent the d-th dimensional coefficient of the source mceps and the target mceps, respectively.
2) Construct the aligned joint feature vectors with the estimated time-warping function.
3) Train a GMM with the joint feature vectors.
4) Convert the source mceps with the trained GMM.
5) Go to step 1 and replace the source mceps with the converted mceps.

The time-alignment function is refined in each iteration because the converted mceps have the same temporal structure as the source mceps but with a more similar speaker individuality to the target speaker. After the process is completed, the resulting time-alignment function is used to construct not only the mceps but also other acoustic features.

III. Time alignment with lip images

Due to the artificial speech generation process, the characteristics of EL speech are different from that of the natural speech, thus the DTW process based on the MCD between mceps can be inaccurate, misleading the estimation of the conversion model. In this work, we consider a scenario where the frontal face video is also recorded when collecting the training data of the EL speaker. We propose to utilize such

[1] https://github.com/k2kobayashi/sprocket

video signals as the input of the DTW-based time-alignment process described in Section II-B.

An essential question to ask is how to choose a proper representation and the corresponding distance measure for the DTW process. Since our goal is to reflect the underlying spoken contents of the video, we hypothesize that the lip images contain the most essential information. In the following subsections, we describe two naive approaches to extract the lip representations from the face video, and the corresponding design choice of the distance metric. Note that other settings of the iterative alignment process remain the same.

A. DTW based on raw lip images

Given a frontal face image as input, the dlib library [16] is used to perform face detection, which is based on a combination of histogram of oriented gradient (HOG) and linear support vector machine (SVM) [17]. Then, the method described in [18] is applied to detect the 68 facial landmarks, including eyes, nose, lips and chin. Based on the 20 landmarks related to the lips, a bounding box can be constructed to extract the raw lip image.

We consider a very simple mean squared error (MSE) between two lip images for the distance metric used in DTW. Since the MSE is calculated in a pixel-wise manner, all lip images are scaled to a predefined size. We denote this approach as DTW-lip-raw.

B. DTW based on lip landmarks

As the mouth positions vary when pronouncing different vowels, such information can be discarded during the scaling step for calculating the pixel-wise MSE in the method described in Section III-A. As a result, two lip images considered close under such representation and distance may not reflect the actual contents, causing errors in the DTW alignment process.

To overcome this problem, we propose to use the landmarks instead of the raw pixels to represent the lips. Specifically, we take the 20 lip landmarks and relocate to the centroid of them. Then, given two sets of lip landmarks from the source and target, denoted as $\mathbf{L}^s = \{(x_1^s, y_1^s), \ldots, (x_{20}^s, y_{20}^s)\}, \mathbf{L}^t = \{(x_1^t, y_1^t), \ldots, (x_{20}^t, y_{20}^t)\}$, we define the following metric:

$$\text{Distance} = \sum_{i=1}^{20} \sqrt{(x_i^s - x_i^t)^2 + (y_i^s - y_i^t)^2}, \qquad (2)$$

which is the sum of the Euclidean distance between each pair of landmarks. By avoiding the scaling process, we believe the alignment process can be more accurate. We refer this method as DTW-lip-landmark.

IV. Experimental evaluations

A. Experimental settings

Experiments were conducted on a Mandarin parallel ELVC corpus. Both the audio and video signals of a doctor who was familiar with the EL device reading the TMHINT dataset [19] with or without the EL device were recorded with a Sony ZV-1. The TMHINT dataset was designed to be phonetically-balanced, where each sentence contained 10 Mandarin Chinese characters. After data cleaning, there were 228 utterances for training and 18 utterances for testing. All speech utterances were sampled at 16 kHz, and the video was recorded in a resolution of 1920×1080 with a frame rate of 50 FPS.

We used the WORLD vocoder [14] for both feature extraction and waveform synthesis. 0-24th mceps were used as the spectral features, and a log-scaled F0, the U/V symbol, and the 513-dimensional aperiodic components were also extracted. The frame shift was set to 5 ms. We downsampled the video stream to 20 FPS, so one lip image corresponds to 4 acoustic frames. Note that to tackle this frame rate mismatch, we stacked 4 acoustic frames to form one long feature vector, such that the alignment path obtained from lips could be directly used.

For the conversion model, we followed the CLDNN [20] structure proposed in [4], and followed most of the settings except for the followings. First, the batch size was set to 16 utterances with zero-padding. The learning rate of all models were set to 0.0005, and the Adam optimizer was used [21].

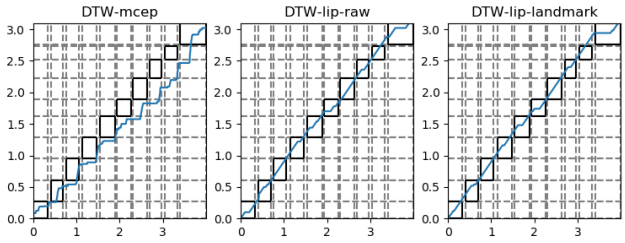

Fig. 4: An example of the alignment matrices obtained with different alignment methods from a parallel EL and normal speech. The blue lines denote the alignment paths, and the dashed grey lines denote the human labeled character-level boundaries.

TABLE I: Comparison of different alignment methods in terms of the correct ratio, which is defined as the overlap rate of the alignment path and the human labeled regions.

Method	DTW-mcep	DTW-lip-raw	DTW-lip-landmark
Ratio	39.73%	46.08%	45.03%

B. Alignment path comparison

Figure 4 visualizes the alignment paths obtained using different alignment methods. To get a sense of the accuracy of the alignments, human-labeled syllable-level boundaries of the EL and normal speech were served as the ground truth and plotted in the figure. Our assumption is that the more overlap between the alignment path and the ground truth region, the better the alignment method. From Figure 4, we observed that the alignment path obtained from DTW-mcep often fell out of the human-labeled regions, while the paths from DTW-lip-raw and DTW-lip-landmark overlapped more with the ground truth boundaries. We also calculated the *correct ratio*, which

TABLE II: Objective evaluation results of models trained with different alignment methods.

Method	Before vocoder		After vocoder	
	MCD	F0RMSE	MCD	F0RMSE
DTW-mcep	7.02	15.84	8.48	28.36
DTW-lip-raw	6.99	14.92	8.41	26.89
DTW-lip-landmark	6.63	13.92	8.09	26.34
Seq2seq [12]	-	-	7.01	26.32

TABLE III: Naturalness (nat.) and intelligibility (int.) preference scores with p-values calculated using a t-test.

Aspect	DTW-mcep	DTW-lip-raw	DTW-lip-landmark	p-value
Nat.	58.1%	41.9%	-	0.014
	42.9%	-	57.1%	0.027
	-	26.3%	73.7%	< 0.001
Int.	42.8%	57.2%	-	0.031
	26.7%	-	73.3%	< 0.001
	-	31.1%	68.9%	< 0.001

is defined as the rate that the alignment path that falls in the regions defined by the human-labeled boundaries. As shown in Table I, the correct ratio using lip-based methods are higher than that of DTW-mcep. These analysis on the alignment paths justify the use of lip images in alignment.

C. Objective evaluation

We carried out two types of objective evaluation. First, the MCD with the same settings described in Section II-B is used since it is a commonly used measure of spectral distortion in VC. We also measured the F0 root mean squared error (F0RMSE), which was calculated using the converted F0 and the target F0. Both the MCD and F0RMSE were calculated in an utterance-wise manner, so DTW was first performed to align the non-silent converted and target mcep sequences beforehand.

As a reference, we included the results of a state-of-the-art seq2seq ELVC model [12]. A Transformer [22] backbone was adopted, and a TTS pretraining strategy [23], [24] was further performed, where the model was pretrained on a large-scale multi-speaker Mandarin TTS dataset followed by fine-tuning on the same Mandarin parallel ELVC corpus. The Parallel WaveGAN (PWG) [25] was chosen as the neural vocoder, which was trained on the training set of the normal speech. For the CLDNN-based methods that used the WORLD features, we reported objective scores both before and after vocoder synthesis. This is because the mcep model and the F0-related models are optimized separately, so the scores after vocoder can reflect the performance of the final generated waveform. For the seq2seq model, since the mel-spectrogram was used as the acoustic feature, only the scores after vocoder synthesis was reported.

The objective evaluation results are shown in Table II. First, by using a simple lip representation and distance, DTW-lip-raw could already outperform DTW-mcep in both metrics. A bigger improvement brought by DTW-lip-landmark showed that a properly designed representation and distance are crucial when using lip images in the temporal alignment process. Finally, it could be clearly observed that there existed a large gap between the MCD values of the DTW-lip-landmark system and the seq2seq model, showing that there is still room for improvements.

D. Subjective evaluation

Finally, we conducted AB tests to assess the subjective preference of models trained with different alignment methods. We measured two different aspects that are important in EL speech enhancement, namely naturalness and intelligibility. To generate the converted speech samples for the listening test, the global variance (GV) post-filter [26] was applied to the converted mceps. We further trained a PWG on the WORLD features extracted from the training normal utterances to generate better sounding samples. We recruited more than 10 native Mandarin speakers as participants. Audio samples can be found online[2].

Table III shows the subjective evaluation results. Compared with DTW-mcep, DTW-lip-raw was inferior in terms of naturalness but superior in intelligibility, while DTW-lip-landmark outperformed in both aspects. When comparing DTW-lip-raw and DTW-lip-landmark, the latter outperformed the former in both naturalness and intelligibility. We conclude that using lip images for alignment can improve the intelligibility of the final VC models, and an improper representation and distance design like DTW-lip-raw can lead to degradation in naturalness. These trends are consistent with the findings in Section IV-C.

V. Conclusions

In this work, we proposed to use lip images to improve temporal alignment in frame-based ELVC, under the assumption that the lip movements are less influenced by the laryngectomy surgery. Two lip representations and distance metrics were investigated, and experimental evaluations were conducted on a Mandarin parallel ELVC corpus both objectively and subjectively. It was demonstrated that using lip images can greatly improve the performance over alignments obtained with acoustic features, and a properly design can lead to a further significant performance gain. For future work, we enumerate several possible improving directions.

Using both acoustic features and lip images in alignment. As mentioned in Section I, the mapping from lip shapes to phonemes is one-to-many, thus solely relying on lip images to perform the alignment may be problematic. On the other hand, some acoustic feature frames that are less affected by the adverse effect of the EL device can be useful in the temporal alignment process. It is therefore worthwhile to investigate using both acoustic features and lip images.

Deep lip representation learning. In this work, we investigated raw lip image pixels and hand-crafted features as the lip representations. An alternative is to use deep feature

[2]https://bit.ly/36aPIpi

representations from a pretrained neural network model. Since the goal of the alignment process is to synchronize according to the underlying phonetic contents, which is considered to be discrete, discrete representation learning models like the vector-quantized variational autoencoder (VQVAE) can be applied.

ACKNOWLEDGMENT

This work was partly supported by JSPS KAKENHI Grant Number 21J20920 and JST CREST Grant Number JPMJCR19A3, Japan. This work was also partly supported by MOST-Taiwan Grant 107-2221-E-001-008-MY3. In addition, this study was approved by a local Institutional Review Board (TMU-JIRB 202005100). Informed consent was obtained from all participants prior to the experiment.

REFERENCES

[1] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech," *Speech Communication*, vol. 54, no. 1, pp. 134–146, 2012.

[2] H. Doi, K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "An evaluation of alaryngeal speech enhancement methods based on voice conversion techniques," in *Proc. ICASSP*, 2011, pp. 5136–5139.

[3] H. Doi, T. Toda, K. Nakamura, H. Saruwatari, and K. Shikano, "Alaryngeal Speech Enhancement Based on One-to-Many Eigenvoice Conversion," *IEEE/ACM TASLP*, vol. 22, no. 1, pp. 172–183, 2014.

[4] K Kobayashi and T Toda, "Electrolaryngeal Speech Enhancement with Statistical Voice Conversion based on CLDNN," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2018, pp. 2115–2119.

[5] T. Dinh, A. Kain, R. Samlan, B. Cao, and J. Wang, "Increasing the Intelligibility and Naturalness of Alaryngeal Speech Using Voice Conversion and Synthetic Fundamental Frequency," in *Proc. Interspeech*, 2020, pp. 4781–4785.

[6] K. Kobayashi and T. Toda, "Implementation of low-latency electrolaryngeal speech enhancement based on multi-task CLDNN," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2021, pp. 396–400.

[7] T. Toda, A. W. Black, and K. Tokuda, "Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory," *IEEE/ACM TASLP*, vol. 15, no. 8, pp. 2222–2235, 2007.

[8] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *Proc. ICASSP*, 1988, vol. 1, pp. 655–658.

[9] A. Kain and M.W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proc. ICASSP*, 1998, pp. 285–288.

[10] L. M Arslan and D. Talkin, "Voice conversion by codebook mapping of line spectral frequencies and excitation spectrum," in *Proc. EUROSPEECH*, 1997, pp. 1347–1350.

[11] Y. Nankaku, K. Nakamura, T. Toda, and K. Tokuda, "Spectral conversion based on statistical models including time-sequence matching," in *Proc. SSW6*, 2007, pp. 333–338.

[12] M.-C. Yen, W.-C. Huang, K. Kobayashi, Y.-H. Peng, S.-W. Tsai, Y. Tsao, T. Toda, J.-S. Roger Jang, and H.-M. Wang, "Mandarin Electrolaryngeal Speech Voice Conversion with Sequence-to-Sequence Modeling," in *Proc. ASRU*.

[13] D. A Ebert and P. S Heckerling, "Communication with deaf patients: knowledge, beliefs, and practices of physicians," *JAMA*, vol. 273, no. 3, pp. 227–229, 1995.

[14] 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. 99, pp. 1877–1884, 2016.

[15] K. Kobayashi and T. Toda, "sprocket: Open-Source Voice Conversion Software," in *Proc. Odyssey*, 2018, pp. 203–210.

[16] D. E. King, "Dlib-ml: A Machine Learning Toolkit," *JMLR*, vol. 10, pp. 1755–1758, 2009.

[17] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. CVPR*, 2005, pp. 886–893.

[18] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proc. CVPR*, 2014, pp. 1867–1874.

[19] M.-W. Huang, "Development of Taiwan Mandarin Hearing In Noise Test," M.S. thesis, Department of speech language pathology and audiology, National Taipei University of Nursing and Health Science, 2005.

[20] T. N Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *Proc. ICASSP*, 2015, pp. 4580–4584.

[21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N Gomez, Ł . Kaiser, and I. Polosukhin, "Attention is All you Need," in *Proc. NIPS*, pp. 5998–6008. 2017.

[23] W.-C. Huang, T. Hayashi, Y.-C. Wu, H. Kameoka, and T. Toda, "Voice transformer network: Sequence-to-sequence voice conversion using transformer with text-to-speech pretraining," in *Proc. Interspeech*, 2020, pp. 4676–4680.

[24] W. C. Huang, T. Hayashi, Y. C. Wu, H. Kameoka, and T. Toda, "Pretraining techniques for sequence-to-sequence voice conversion," *IEEE/ACM TASLP*, vol. 29, pp. 745–755, 2021.

[25] R. Yamamoto, E. Song, and J. Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *Proc. ICASSP*, 2020, pp. 6199–6203.

[26] H. Silén, E. Helander, J. Nurminen, and M. Gabbouj, "Ways to implement global variance in statistical speech synthesis," in *Proc. Interspeech*, 2012, pp. 1436–1439.