15th IAPR International Conference on Machine Vision Applications (MVA)
Nagoya University, Nagoya, Japan, May 8-12, 2017.

Visual-to-Speech Conversion Based on Maximum Likelihood Estimation

Rina Ra, Ryo Aihara, Tetsuya Takiguchi, Yasuo Ariki
Graduate School of System Informatics, Kobe University
1-1, Rokkodai, Nada, Kobe, Japan
{rinara, aihara}@me.cs.scitec.kobe-u.ac.jp, {takigu, ariki}@kobe-u.ac.jp

Abstract

This paper proposes a visual-to-speech conversion method that converts voiceless lip movements into voiced utterances without recognizing text information. Inspired by a Gaussian Mixture Model (GMM)-based voice conversion method, GMM is estimated from jointed visual and audio features and input visual features are converted to audio features using maximum likelihood estimation. In order to capture lip movements whose frame rate data is smaller than the audio data, we construct long-term image features. The proposed method has been evaluated using large-vocabulary continuous speech and experimental results show that our proposed method effectively estimates spectral envelopes and fundamental frequencies of audio speech from voiceless lip movements.

1 Introduction

Visual-to-Speech Conversion (VTSC) is a technique that converts "unvoiced" lip movements to "voiced" utterances. McGurk *et al.* [1] reported that we perceive a phoneme not only from voice included auditory information, but also from visual information gathered from the speaker's lips and facial movements. Moreover, it has been reported that we try to catch the movement of lips in a noisy environment and we misunderstand the utterance when the movements of the lips and the voice are not synchronized. VTSC is a difficult challenge because visual images contain less linguistic information than audio speech; however, we assume VTSC will be an assistive technology for those who have a speech impediment or that it can be adopted to voice reconstruction of videos lacking sound tracks or communication tools in noisy environments.

There are two approaches to obtain audio speech from voiceless lip movements. The first approach is a combined method of lip reading and Text-To-Speech (TTS) synthesis. Lip reading is a technique that recognizes text information from voiceless lip movements. In this approach, input lip movement is recognized using lip reading and the estimated text is synthesized to target voice utterances using TTS systems. This approach can be called "visual-to-speech synthesis". The other approach is a more direct one that does not recognize the text information of input lip movements. The former approach may be effective because of recent developments in lip reading [2] and TTS [3]. However, the linguistic information of the output voice will be incorrect when the lipreading system failed to recognize the text. Moreover, the first method needs a large amount of training data to develop the lipreading and TTS system. Therefore, this paper adopts the latter approach, and we call this approach VTSC.

In the field of speech-signal processing, there are techniques similar to VTSC. Voice Conversion (VC) converts para-linguistic information such as speaker identification while maintaining other information, such as linguistic information, in the speech utterance. A number of VC methods have been proposed [4, 5, 6, 7], and most of them do not recognize the text information from the input utterance. The Gaussian Mixture Model (GMM)-based approach is widely used for VC because of its flexibility and good performance [4]. In this approach, source and target spectral features are approximated by GMM and the conversion function is interpreted as the expectation value of the target spectral envelope. The conversion parameters are evaluated using minimum mean-square error or maximum likelihood (ML) on a training set [8].

Inspired by the GMM-based VC method [8], we propose a novel VTSC method based on ML estimation. Visual features and audio features are jointed and they are approximated by GMM. Input visual features are converted to audio features by using ML estimation. In the case of VC, short-term spectral features are used; however, it is not effective for VTSC because the frame rate of the visual data is smaller than audio data and visual features contain less information compared to audio features. Therefore, we construct a long-term image feature, which contains multiple frames of images. We estimate the spectral envelope and Fundamental frequency (F0) from visual image; however, these two features are estimated independently. Experimental results show that our proposed VTSC can effectively estimate the spectral envelope and F0 from input lip movements of large-vocabulary continuous speech.

There are some related works. Speech-to-lip movement synthesis is an inverse problem to VTSC. A recognition-based approach using hidden Markov models has been widely researched [9]. Lavagetto [10] applied neural networks to speech-to-lip conversion for assistive technology for the people with hearing loss. Zhuang, *et al.* [11] applied a GMM-VC method [8] to speech-to-lip conversion. Lip-to-speech synthesis using non-negative matrix factorization has also proposed [12]; however, it works only for limited data, such as digit utterances.

The rest of this paper is organized as follows: In Section 2, our proposed method is described. In Section 3, the experimental data are evaluated, and the final section is devoted to our conclusions.

2 Proposed Visual-to-Speech Conversion

2.1 Feature construction

Fig. 1 shows the flow chart of the visual feature extraction. First, Region of Interest (ROI) is extracted from visual images. The brightness of the image is regularized so that they have flat frequency distributions. Then, 2-dimensional Discrete Cosine Transform (2D-DCT) is applied to the image, and a zigzag scan is used to obtain the 1D-DCT coefficient vector. Obtained coefficient vectors are normalized by Z-score. In order to fill the sampling rate gap between audio features, visual features are interpolated by spline interpolation, and static image features are obtained.

In order to capture the lip movements, we construct long-term image features. Fig. 2 shows the flow of the construction. $d_x(2L+1)$-dimensional segmental features are constructed from the d_x-dimensional static image vectors $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$, where T denotes the number of the frame. Principal Component Analysis (PCA) is applied to the segmental feature and D_x dimensional long-term image vectors $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_T\}$ are constructed.

For the audio features, spectral envelope, F0, and aperiodic components are extracted by using a vocoder named STRAIGHT [13]. In this paper, the spectral envelope and F0 are independently estimated from visual features and aperiodic components are not considered. For spectrum estimation, the dimensional mel-cepstrum d_y is calculated from STRAIGHT spectrum, and its delta features are jointed to the static mel-cepstrum. Static mel-cepstrum $\mathbf{y}$ and its delta features $\Delta\mathbf{y}$ are used as target audio feature vectors i.e. $\mathbf{Y} = [\mathbf{y}^\mathsf{T} \Delta\mathbf{y}^\mathsf{T}]^\mathsf{T}$. For F0 estimation, log-scaled F0 and delta features are used as $\mathbf{Y}$.

We can also use long-term features for audio features by adopting the flow in Fig. 2. However, in order to estimate continuous audio features in the conversion stage, we use a trajectory model, which considers the relationship between target static features and its delta features.

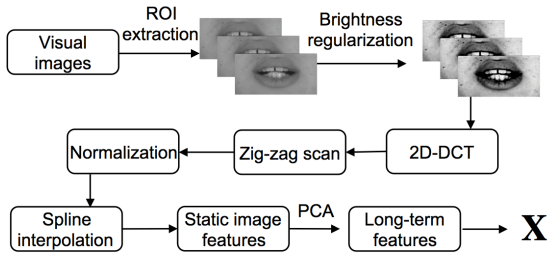

Figure 1. Flow of the visual feature extraction.

2.2 Maximum likelihood-based conversion

We model a joint probability of image features and audio features using the mixture of multivariate Gaussian distribution $\mathcal{N}(.; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ with parameters of a mean vector $\boldsymbol{\mu}$ and a variance matrix $\boldsymbol{\Sigma}$. Therefore, this model is called a "joint density GMM" (JD-GMM). In the training stage of the JD-GMM, we use a joint vector $\mathbf{Z}$ that concatenates image feature vector $\mathbf{X}$ and audio feature vector $\mathbf{Y}$ (i.e. $\mathbf{Z} = [\mathbf{X}^\mathsf{T}\mathbf{Y}^\mathsf{T}]^\mathsf{T}$). The probability $p(\mathbf{Z})$ is modeled using GMM as follows:

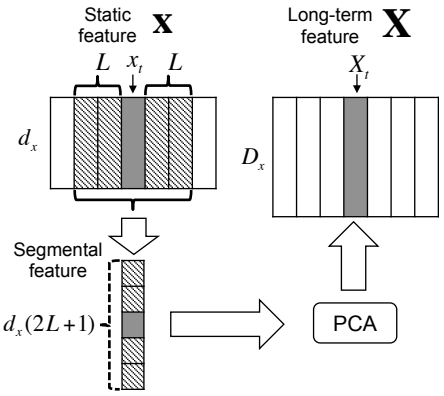

Figure 2. Flow of the construction of long-term image features.

$$p(\mathbf{Z}|\boldsymbol{\Theta}^{(z)}) = \sum_{m=1}^{M} \alpha_m \mathcal{N}(\mathbf{Z}; \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)}), \tag{1}$$

where $\boldsymbol{\mu}_m^{(z)}$ and $\boldsymbol{\Sigma}_m^{(z)}$ consist of

$$\boldsymbol{\mu}_m^{(z)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(x)} \\ \boldsymbol{\mu}_m^{(y)} \end{bmatrix}, \quad \boldsymbol{\Sigma}_m^{(z)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(xx)} & \boldsymbol{\Sigma}_m^{(xy)} \\ \boldsymbol{\Sigma}_m^{(yx)} & \boldsymbol{\Sigma}_m^{(yy)} \end{bmatrix}. \tag{2}$$

The parameters $\boldsymbol{\mu}_m^{(x)}$ and $\boldsymbol{\Sigma}_m^{(xx)}$, and the parameters $\boldsymbol{\mu}_m^{(y)}$ and $\boldsymbol{\Sigma}_m^{(yy)}$ correspond to the Gaussian distribution of image features and Gaussian distribution of audio features, respectively. α_m denotes the weight of m-th component. The parameter $\boldsymbol{\Sigma}_m^{(xy)} (= \boldsymbol{\Sigma}_m^{(yx)^\mathsf{T}})$ indicates a covariance matrix between the observed data $\mathbf{X}$ and $\mathbf{Y}$. $\boldsymbol{\Theta}^z$ is a set of parameters of GMM, which contains α_m, $\boldsymbol{\mu}_m^{(x)}$, $\boldsymbol{\mu}_m^{(y)}$, $\boldsymbol{\Sigma}_m^{(xx)}$, $\boldsymbol{\Sigma}_m^{(yy)}$, and $\boldsymbol{\Sigma}_m^{(xy)}$ for all m. M denotes the number of Gaussian mixtures.

In VC, we usually use a diagonal matrix for $\boldsymbol{\Sigma}_m^{(xx)}$, $\boldsymbol{\Sigma}_m^{(xy)}$, and $\boldsymbol{\Sigma}_m^{(yy)}$ to reduce the number of parameters. Such parameter reduction is effective in VC because both $\mathbf{X}$ and $\mathbf{Y}$ are the same acoustic features. However, in VTSC, we use full-covariance matrices because $\mathbf{X}$ and $\mathbf{Y}$ are different features. The GMM parameters can be estimated using the Expectation-Maximization (EM) algorithm.

In the conversion stage, we consider the probability of $\mathbf{Y}$ given an input $\mathbf{X}$. That is

$$\begin{aligned} & p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\Theta}^{(z)}) \\ &= \sum_{all\,\mathbf{m}} p(\mathbf{m}|\mathbf{X}, \boldsymbol{\Theta}^{(z)}) p(\mathbf{Y}|\mathbf{X}, \mathbf{m}, \boldsymbol{\Theta}^{(z)}) \\ &= \prod_{t=1}^{T} \sum_{m_t=1}^{M} p(m_t|\mathbf{X}_t, \boldsymbol{\Theta}^{(z)}) p(\mathbf{Y}_t|\mathbf{X}_t, m_t, \boldsymbol{\Theta}^{(z)}) \end{aligned} \tag{3}$$

where $\mathbf{m} = \{m_1, m_2, \cdots, m_T\}$ is a mixture component sequence. The probabilities on the right side in Eq. (3) are represented as

$$p(m_t|\mathbf{X}_t, \boldsymbol{\Theta}^{(z)}) = \frac{\alpha_m \mathcal{N}(\mathbf{X}_t; \boldsymbol{\mu}_m^{(x)}, \boldsymbol{\Sigma}_m^{(xx)})}{\sum_{n=1}^{M} \alpha_n \mathcal{N}(\mathbf{X}_t; \boldsymbol{\mu}_n^{(x)}, \boldsymbol{\Sigma}_n^{(xx)})} \tag{4}$$

$$p(\mathbf{Y}_t|\mathbf{X}_t, m_t, \boldsymbol{\Theta}^{(z)}) = \mathcal{N}(\mathbf{Y}_t; \mathbf{E}_{m,t}^{(y|x)}, \mathbf{D}_m^{(y|x)}) \tag{5}$$

where

$$\mathbf{E}_{m,t}^{(y|x)} = \boldsymbol{\mu}_m^{(y)} + \boldsymbol{\Sigma}_m^{(yx)}(\boldsymbol{\Sigma}_m^{(xx)})^{-1}(\mathbf{X}_t - \boldsymbol{\mu}_m^{(x)}) \quad (6)$$

$$\mathbf{D}_m^{(y|x)} = \boldsymbol{\Sigma}_m^{(yy)} - \boldsymbol{\Sigma}_m^{(yx)}(\boldsymbol{\Sigma}_m^{(xx)})^{-1}\boldsymbol{\Sigma}_m^{(xy)}. \quad (7)$$

A time sequence of the converted feature vector $\hat{\mathbf{y}}$ is determined as follows:

$$\hat{\mathbf{y}} = \arg\max P(\mathbf{Y}|\mathbf{X}, \boldsymbol{\Theta}^{(z)}). \quad (8)$$

Eq. (8) is performed under the linear conversion between static feature vectors $\mathbf{y}$ and the static and dynamic feature vectors $\mathbf{Y}$:

$$\mathbf{Y} = \mathbf{W}\mathbf{y} \quad (9)$$

where $\mathbf{W}$ is a transformation matrix [8].

Eq. (3) are approximated with a single mixture component sequence as follows:

$$p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\Theta}^{(z)}) \simeq p(\hat{\mathbf{m}}|\mathbf{X}, \boldsymbol{\Theta}^{(z)})p(\mathbf{Y}|\mathbf{X}, \hat{\mathbf{m}}, \boldsymbol{\Theta}^{(z)}). \quad (10)$$

$\hat{\mathbf{m}}$ denotes the suboptimum mixture component sequence which is determined as follows:

$$\hat{\mathbf{m}} = \arg\max P(\mathbf{m}|\mathbf{X}, \boldsymbol{\Theta}^{(z)}). \quad (11)$$

The logarithm of the likelihood function is written as

$$\begin{aligned}&\log p(\mathbf{Y}|\mathbf{X}, \hat{\mathbf{m}}, \boldsymbol{\Theta}^{(z)}) \\ &= -\frac{1}{2}\mathbf{Y}^\top \mathbf{D}_{\hat{\mathbf{m}}}^{(y|x)^{-1}}\mathbf{Y} + \mathbf{Y}^\top \mathbf{D}_{\hat{\mathbf{m}}}^{(y|x)^{-1}}\mathbf{E}_{\hat{\mathbf{m}}}^{(y|x)} + K \quad (12)\end{aligned}$$

where

$$\mathbf{E}_{\hat{\mathbf{m}}}^{(y|x)} = [\mathbf{E}_{\hat{m}_1,1}^{(y|x)}, \mathbf{E}_{\hat{m}_2,2}^{(y|x)}, \cdots, \mathbf{E}_{\hat{m}_T,T}^{(y|x)}] \quad (13)$$

$$\mathbf{D}_{\hat{\mathbf{m}}}^{(y|x)} = \mathrm{diag}[\mathbf{D}_{\hat{m}_1,1}^{(y|x)}, \mathbf{D}_{\hat{m}_2,2}^{(y|x)}, \cdots, \mathbf{D}_{\hat{m}_T,T}^{(y|x)}]. \quad (14)$$

we can estimate the most probable $\hat{\mathbf{y}}$ as follows:

$$\hat{\mathbf{y}} = (\mathbf{W}^\top \mathbf{D}_{\hat{\mathbf{m}}}^{(y|x)^{-1}}\mathbf{W})^{-1}\mathbf{W}^\top \mathbf{D}_{\hat{\mathbf{m}}}^{(y|x)^{-1}}\mathbf{E}_{\hat{\mathbf{m}}}^{(y|x)}. \quad (15)$$

We can also maximize the logarithm of the likelihood function of Eq. (3) by employing the EM algorithm. However, in VC, there is little difference between the conversion accuracy when using the suboptimum mixture component sequence and the conversion accuracy when using the EM algorithm [8] therefore we also adopt the conversion using the suboptimum mixture component sequence to VTSC.

3 Experimental Results

3.1 Experimental conditions

The proposed VTSC was evaluated on the M2TINIT database [14], which contains audio and visual images of utterances spoken by one Japanese male. Five-hundred and three Japanese-phoneme balanced sentences are included. The number of the training sentences was chosen from the set $\{50, 100, 200\}$. Fifty sentences, which were not included in the test data, were used for testing.

The video images contained only images of the area from the mouth to the tip of the nose. The original frame rate of images is 1/29.97 sec., and they were interpolated so that they have the same frame rate as the audio features. The size of the image was 720×480-pixels, and a 40×20-pixels mouth area was extracted. We introduced 25-dimensional DCT features as static image features, and set the number of dimensions of the long-term image feature at 50.

The sampling frequency of the audio speech data was 16kHz, and the frame shift was 5ms. Each sample was analyzed by STRAIGHT [13], and F0, spectral envelope, and aperiodic components were extracted. Mel-cepstral features, which are used as spectral features, were calculated from the STRAIGHT spectral envelope, and Δ features were added to them. The energy of the mel-cepstrum was also used, and the total number of dimensions of spectral features was 50. Mel-cepstrum Distortion (MelCD) [dB] was used as a measure of the objective evaluations of spectrum estimation which is defined as follows:

$$MelCD = (10/\log 10)\sqrt{2\sum_{d}^{24}(mc_d^{conv} - mc_d^{tar})^2} \quad (16)$$

where mc_d^{conv} and mc_d^{tar} denote the d-th dimension of the converted and target mel-cepstra, respectively.

For F0 estimation, the number of dimensions of F0 features was 2, which contains static and delta features. The estimated spectral envelope and F0 were synthesized to speech signals using STRAIGHT, where aperiodic components were not considered. The number of Gaussians was chosen from the set $\{8, 16, 32, 64, 128, 256\}$. We used Root Mean Square Error (RSME) as a measure of the objective evaluations of log-F0 estimation.

3.2 Results and discussion

First, we evaluated the effectiveness of the long-term image feature for spectrum estimation. Fig. 3 shows the MelCD using different image features. Static+delta denotes the joint feature of static DCT and its delta feature. PCA denotes the long-term image feature and, L is explained in Fig. 2. As shown in the figure, the long-term image feature using $L = 3$ is the most effective for spectrum estimation.

Next, we evaluated the spectrum estimation using the different number of training sentences and the results are shown in Fig. 4. The distortion decreases as we increase the number of training sentences.

Finally, we evaluated the F0 estimation using the different image features. Fig. 5 shows that the long-term image features are also effective for F0 estimation.

Fig. 6 and 7 show examples of spectrogram of the target audio and the estimated audio.

4 Conclusions

This paper proposed a statistical technique for spectrum and F0 estimation from image features. We defined VTSC (Visual-to-Speech Conversion) enables the reconstruction of voiced utterances from unvoiced lip movement images without recognizing text information. Spectrum envelopes or F0 are jointed with image features, and GMM independently models them. Target acoustic features are obtained by using ML estimation. In order to capture the movement of the lip from

image data whose frame rate data is smaller than the audio data, we used long-term image features, which considered multiple image frames. Objective evaluations showed that our proposed VTSC method effectively estimated the acoustic features from image features. Our future work includes the evaluation of the other advanced image features and increasing the number of the subjects.

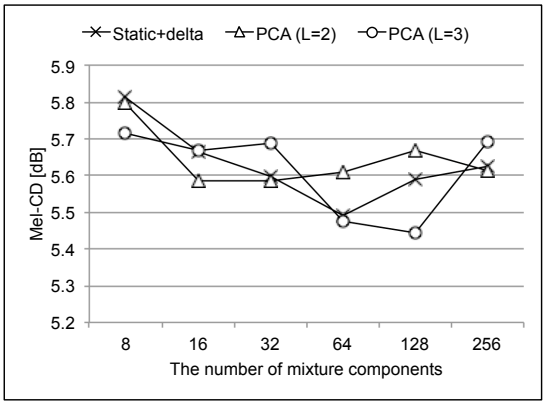

Figure 3. MelCD as a function of the number of mixture components using 100 training sentences.

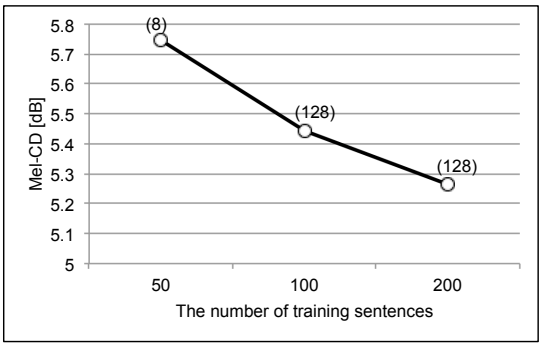

Figure 4. MelCD as a function of the number of training sentences using PCA ($L = 3$). The numbers within parentheses indicate the optimum number of mixture components.

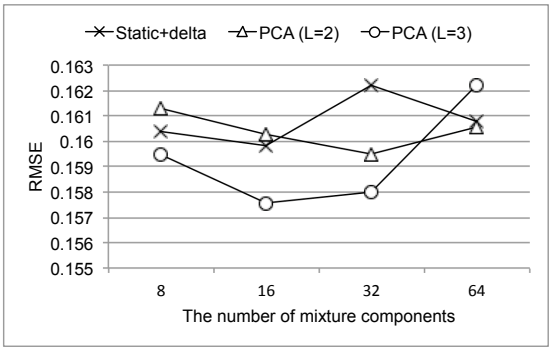

Figure 5. RMSE as a function of the number of mixture components using 200 training sentences.

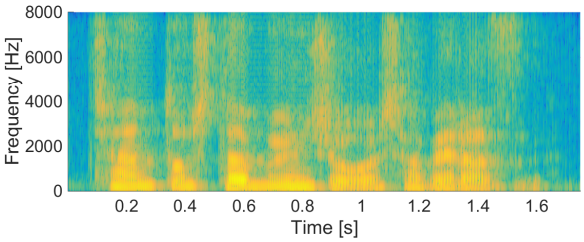

Figure 6. An example audio target spectrogram.

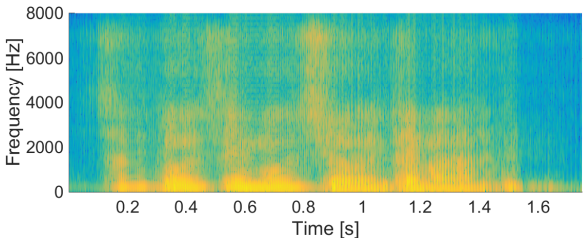

Figure 7. An example of an estimated audio spectrogram using PCA ($L = 3$) and 200 training sentences.

References

[1] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.

[2] Y. M. Assael *et al.*, "Lipnet: Sentence-level lipreading," *arXiv:1611.01599*, 2016.

[3] A. van den Oord *et al.*, "Wavenet: A generative model for raw audio," *CoRR*, vol. abs/1609.03499, 2016.

[4] Y. Stylianou *et al.*, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.

[5] C. Ling-Hui *et al.*, "Joint spectral distribution modeling using restricted boltzmann machines for voice conversion," *in Proc. Interspeech*, pp. 3052—3056, 2013.

[6] R. Aihara *et al.*, "Multiple non-negative matrix factorization for many-to-many voice conversion," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 24, no. 7, pp. 1175–1184, 2016.

[7] K. Nakamura *et al.*, "Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech," *Speech Communication*, vol. 54, no. 1, pp. 134–146, 2012.

[8] T. Toda *et al.*, "Voice conversion based on maximum likelihood estimation of spectral parameter trajectory," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2222–2235, 2007.

[9] E. Yamamoto *et al.*, "Lip movement synthesis from speech based on Hidden Markov Models," *Speech Communication*, vol. 25, no. 1-2, pp. 105–115, 1998.

[10] F. Lavagetto, "Converting speech into lip movements: a multimedia telephone for hard of hearing people," *IEEE Trans. on Rehabilitation Engineering*, vol. 3, no. 1, pp. 90–102, 1995.

[11] X. Zhuang *et al.*, "A minimum converted trajectory error (MCTE) approach to high quality speech-to-lips conversion," *in Proc. INTERSPEECH*, pp. 1736–1739, 2010.

[12] R. Aihara *et al.*, "Lip-to-speech synthesis using locality-constraint non-negative matrix factorization," *in Proc. MLSLP*, 2015.

[13] H. Kawahara, "STRAIGHT, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds," *Acoustical Science and Technology*, pp. 349–353, 2006.

[14] S. Sako *et al.*, "HMM-based text-to-audio-visual speech synthesis –image-based approach," *in Proc. ICSLP*, vol. 3, pp. 25–28, 2000.