Conversion of Lip Movements into Speech Using Gaussian Mixture Models

Rina Ra, Ryo Aihara, Tetsuya Takiguchi, and Yasuo Ariki

***Abstract*— This paper describes a novel lip-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, a GMM is estimated from jointed lip-movements and audio features, and for test, an input lip-movements feature is converted to the audio feature using maximum likelihood (ML) estimation. The proposed method has been evaluated using large-vocabulary continuous utterances and experimental results show that our proposed method effectively estimates spectral envelopes and fundamental frequencies of audio speech from voiceless lip movements.**

I. Introduction

Lip-to-Speech Conversion (LTSC) is a technique that converts "unvoiced" lip movements to "voiced" utterances [1][2], and it is a difficult challenge because visual images contain less linguistic information than audio speech. However, we assume LTSC will be an assistive technology for those who have a speech impediment or communication tools in noisy environments.

In this paper, a novel LTSC method based on ML estimation is described. In the training process, visual (lip) features and audio features are jointed, and they are approximated by a GMM. Then, an input visual feature is converted to the audio feature (spectral envelope and fundamental frequency) by using the ML estimation, where a long-term image feature is constructed from multiple frames of images.

II. Lip-to-Speech conversion

In order to capture the lip movements, a segmental image feature is constructed by concatenating the $\pm L$ consecutive frames of the image feature. Then, Principal Component Analysis (PCA) is applied to the segmental feature in order to obtain the long-term image feature.

For the audio features in the training process, spectral envelope, F0 (fundamental frequency), and aperiodic components are extracted by using a vocoder named STRAIGHT [3]. In this paper, the spectral envelope and F0 are independently estimated from visual features, and aperiodic components are not considered. For F0 estimation, log-scaled F0 and delta features are used.

A joint probability of a joint vector Z of image features X and audio features Y is modeled using the mixture of multivariate Gaussian distribution $N(.;\mu,\Sigma)$ with parameters of a mean vector and a variance matrix in the training process. In the conversion process, the probability of Y given an input X is considered, and a time sequence of the converted feature vector is determined using maximum likelihood estimation [2]:

$$\hat{y} = argmax\, P(Y|X,\Theta^{(Z)}) \tag{1}$$

III. Results and discussion

The number of training sentences was 300, and fifty sentences were used for testing. The size of the image was 720×480 pixels, and a 40×20-pixels mouth area was extracted. Fig. 1 shows the effectiveness of the long-term image feature for acoustic spectrum estimation, where mel-cepstrum distortion [dB] was used as a measure of the objective evaluations. As shown in the figure (for the number of image feature dimensions after PCA: 50, 100, and 150), the long-term image feature using L = 10 or 20 is the most effective for acoustic spectrum estimation.

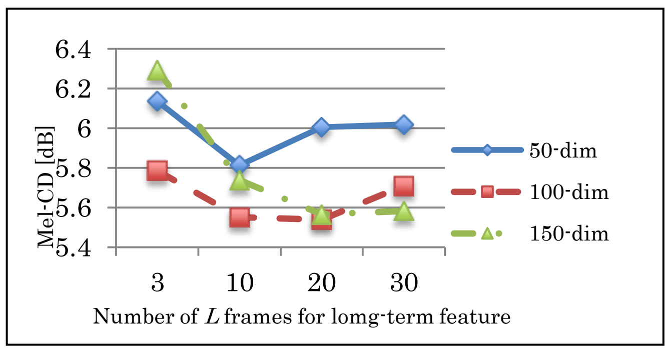

Figure 1. Mel-cep distortion (for the acoustic spectral feature) as a function of the number of dimensions for the long-term image feature.

Our future work includes the evaluation of the other advanced image features and the subjective evaluations.

Acknowledgment

This work was supported in part by PRESTO, JST (Grant No. JPMJPR15D2).

References

[1] Ryo Aihara, *et al.*, "Li-to-speech synthesis using locality-constraint non-negative matrix factorization," in *Proc. of International Workshop on Machine Learning in Spoken Language Processing*, 2015, 6 pages.

[2] Rina Ra, *et al.*, "Visual-to-Speech Conversion Based on Maximum Likelihood Estimation," in *Proc. of International Conference on Machine Vision Applications*, 2017.

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

*This work was supported in part by PRESTO, JST.

Rina Ra, Ryo Aihara, Tetsuya Takiguchi, and Yasuo Ariki are with Kobe University, Japan (e-mail: rinara@me.cs.scitec.kobe-u.ac.jp).