# LIP-TO-SPEECH SYNTHESIS USING LOCALITY-CONSTRAINT NON-NEGATIVE MATRIX FACTORIZATION

*Ryo AIHARA, Kenta MASAKA, Tetsuya TAKIGUCHI, Yasuo ARIKI*

Graduate School of System Informatics, Kobe University, Japan

## ABSTRACT

We propose in this paper a lip-to-speech conversion method that converts "unvoiced" lip movements to "voiced" utterances, where parallel lip movements and speech spectra are stored as a source dictionary and a target dictionary, respectively. An input lip image is decomposed into a linear combination of bases from the source dictionary, and its weight is estimated by Non-negative Matrix Factorization (NMF). The selected image bases are replaced with speech bases from the target dictionary, and the speech spectra are constructed. We assume this method will be an assistive technology for people who have speech disabilities. In this paper, NMF using $\beta$ divergence is used as a cost function and introduced locality-constraint in order to increase sparsity in an activity matrix. The effectiveness of our method was confirmed by objective and subjective evaluations.

***Index Terms***— speech synthesis, multimodal, assistive technology

## 1. INTRODUCTION

An assistive technology is a system or a product which is used to improve the functional capabilities of individuals with disabilities. For past few decades, some speech processing techniques have been adopted to the assistive technology. As a consequent of recent advance in statistical text-to-speech synthesis (TTS), Hidden Markov Model (HMM)-based TTS is used for reconstructing the voice of individuals with degenerative speech disorders [1]. Voice conversion (VC) is also applied to the assistive technology. A Gaussian mixture model (GMM)-based VC method has been applied to reconstruct speaker's individuality in electrolaryngeal speech [2] and speech recorded by Non-Audible Murmur (NAM) microphones [3].

In this paper, we propose a lip-to-speech synthesis using a sparse representation technique. Lip images without a voice recording are converted to a voice utterance. We assume our proposed method will be an assistive technology for those who have a speech impediment. There are 34,000 such people in Japan alone; therefore, there is a great need for such a technology. Moreover, our approach can be adopted to voice reconstruction of videos lacking sound tracks or communication tools in noisy environments.

Lip reading is a technique of understanding speech by visually interpreting the movements of the lips, face and tongue when the spoken sounds cannot be heard. For example, for people with hearing problems, lip reading is one communication skill that can help them communicate better. McGurk *et al.* [4] reported that we perceive a phoneme not only from auditory information from the voice but also from visual information from the lips or from facial movements. Moreover, it is 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. In the field of speech processing, audio-visual speech recognition has been researched for robust speech recognition under noisy environments [5, 6]. However, as far as our knowledge, lip-to-speech synthesis has never been proposed.

We used Non-negative Matrix Factorization (NMF) [7], which is a famous approach using sparse representations. Sparse representations are employed to speech separation [8], super resolution [9], etc. NMF is widely used in the field of speech processing, and we used it in a VC framework [10]. In this approach, parallel dictionaries, which consist of the same utterances of the source speaker and the target speaker, are needed. An input source speaker's utterance is decomposed into a linear combination of a small number of bases from the source dictionary. The selected bases are replaced with the bases of the target dictionary, which are parallel to the source bases. We adopted this method for lip-to-speech synthesis. We need parallel exemplars of visual and audio speech data for system construction; however, in a test phase, a voiceless lip image can be converted to a voiced utterance without utterance recognition techniques such as standard VC.

Our dictionary is over-complete and contains a large number of bases. Because lip movements closely resemble each other compared to speech spectra, lip images may be decomposed into a large number of bases, which can lead to a degradation of the converted sound. Therefore, in this paper, we introduce a locality-constraint to the activities of NMF [11] in order to increase the sparseness.

The rest of this paper is organized as follows: In Section 2, related works are introduced. In Section 3, NMF using $\beta$ divergence is described. In Section 4, our proposed method is explained. In Section 5, the experimental data are evaluated, and the final section is devoted to our conclusions.

## 2. RELATED WORKS

Lee *et al.* [7] proposed an NMF algorithm using a maximization-minimization algorithm. The cost function used in [7] was the Euclidean distance and the Kullback-Leibler (KL) divergence. NMF using the Itakura-Saito (IS) divergence has also been proposed [12] because the IS divergence was presented as "a measure of the goodness of fit between two spectra". Eguchi *et al.* [13] introduced $\beta$ divergence, which is a family of cost functions that includes the Euclidean distance, the KL divergence and the IS divergence. The algorithm of NMF using the $\beta$ divergence is summarized in [14]. In the field of speech processing, NMF has been used for speech separation [8], music transcription [15], noise-robust speech recognition [16], etc. In [10, 17], we proposed VC using NMF, and this proposed method was the inspiration behind our lip-to-speech synthesis approach. NMF-based VC has been also adopted as an assistive technology for articulation disorders [18]. In our previous work [19], we proposed a multimodal NMF-based VC method which converts audio-visual features to target speaker's audio features. In almost all

approaches for speech processing using NMF, the KL divergence or the $\beta$ divergence is used as its cost function. Multimodal statistical VC was proposed in [20, 21]

In the case of image processing, Orthogonal Matching Pursuit (OMP) [22] or Sparse Coding (SC) is the popular algorithm of sparse representation. Those algorithms are used for image clustering [23], super resolution [9] etc. In most cases, the Euclidean distance is used for the cost function of these algorithms. In this paper, we employed NMF using the $\beta$ divergence and searched for the best value of $\beta$.

Audio-visual speech recognition and lipreading have been researched in the field of signal processing. Tao and Busso [24] proposed audio-visual whisper isolated digits recognition using HMMs. Noda *et al.* [25] proposed convolutional neural network-based visual feature extraction for lipreading. However, the word recognition rate of the state-of-the-art lipreading in a closed speaker open-vocabulary task is less than 40%. This result shows the difficulties in lipreading. The other silent speech interfaces, for example NAM microphones, are introduced in [26].

Speech-to-lip movement synthesis is an inverse problem to our lip-to-speech synthesis. Speech-to-lip synthesis has been researched for the needs of avatar talk on the Internet. A recognition-based approach using HMMs has been widely researched [27]. Lavagetto [28] applied neural networks to speech-to-lip conversion for the assistive technology for the people with hearing loss. Zhuang [29] applied a statistical VC method using a Gaussian Mixture Model (GMM) [30] to speech-to-lip conversion, which does not need utterance recognition of lip images.

## 3. NON-NEGATIVE MATRIX FACTORIZATION USING $\beta$ DIVERGENCE

In the approaches based on sparse representations, the observed signal is represented by a linear combination of a small number of bases.

$$\mathbf{x}_l \approx \sum_{j=1}^{J} \mathbf{w}_j h_{j,l} = \mathbf{W}\mathbf{h}_l \tag{1}$$

$\mathbf{x}_l$ represents the $l$-th frame of the observation. $\mathbf{w}_j$ and $h_{j,l}$ represent the $j$-th basis and the weight, respectively. $\mathbf{W} = [\mathbf{w}_1 \dots \mathbf{w}_J]$ and $\mathbf{h}_l = [h_{1,l} \dots h_{J,l}]^{\mathbf{T}}$ represent the collection of the bases and the stack of weights. When the weight vector $\mathbf{h}_l$ is sparse, the observed signal can be represented by a linear combination of a small number of bases that have non-zero weights. In this paper, each basis denotes the exemplar of the speech or image signal, and the collection of exemplar $\mathbf{W}$ and the weight vector $\mathbf{h}_l$ are called the 'dictionary' and 'activity', respectively. When feature vectors are lined up, (1) is represented as an inner product of feature matrix.

$$\mathbf{X} \approx \mathbf{W}\mathbf{H} \tag{2}$$

$$\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_L], \quad \mathbf{H} = [\mathbf{h}_1, \dots, \mathbf{h}_L]. \tag{3}$$

where $L$ denotes the number of frames.

In this paper, we employ NMF in order to estimate an activity matrix. The cost function of NMF is defined as follows:

$$d_\beta(\mathbf{x}_l, \mathbf{W}\mathbf{h}_l) + \lambda||\mathbf{h}_l||_1 \quad s.t. \quad \mathbf{h}_l \geq 0 \tag{4}$$

The first term is the $\beta$ divergence between $\mathbf{x}_l$ and $\mathbf{W}\mathbf{h}_l$. The second term is the sparse constraint with the L1-norm regularization term that causes $\mathbf{h}_l$ to be sparse.

The $\beta$ divergence is parameterized with a parameter $\beta$, which takes the Euclidean distance ($\beta = 2$), the KL divergence ($\beta = 1$) and the IS divergence ($\beta = 0$) as follows:

$$d_\beta(\mathbf{x}, \mathbf{y}) = \begin{cases} \frac{1}{\beta(\beta-1)}(\mathbf{x}^\beta + (\beta-1)\mathbf{y}^\beta - \beta\mathbf{x}\mathbf{y}^\beta) & \beta \in \mathbb{R}\backslash\{0,1\} \\ \mathbf{x}\log(\mathbf{x}/\mathbf{y}) - \mathbf{x} + \mathbf{y} & \beta = 1 \\ (\mathbf{x}/\mathbf{y}) - \log(\mathbf{x}/\mathbf{y}) - 1 & \beta = 0 \end{cases}$$

$\mathbf{h}$ minimizing (4) is estimated iteratively by applying the following update rule [14]:

$$h_l \leftarrow h_l \left( \frac{\sum_j w_{jl} x_l \hat{x}_l^{\beta-2}}{\sum_j w_{jl} \hat{x}_l^{\beta-1}} \right)^{\gamma(\beta)} \tag{5}$$

where we denote $[\mathbf{W}\mathbf{h}]_l = \hat{x}_l$ and $\gamma(\beta)$ as follows:

$$\gamma(\beta) = \begin{cases} \frac{1}{2-\beta} & \beta < 1 \\ 1 & 1 \leq \beta \leq 2 \\ \frac{1}{\beta-1} & \beta > 2 \end{cases} \tag{6}$$

In our proposed method, the dictionary $\mathbf{W}$ is obtained by just lining up the parallel data. Therefore, it does not use any training algorithm to obtain the source dictionary.

## 4. SPEECH PRODUCTION USING NMF

### 4.1. Flow of the Proposed Method

Fig. 1 shows the flow of our proposed method. $\mathbf{X}^V$, $\mathbf{W}^V$, $\mathbf{W}^A$ and $\mathbf{X}^A$ denote input image features ($D_v \times L$), source visual dictionary ($D_v \times J$), target audio dictionary ($D_a \times J$), and produced audio features ($D_a \times L$), respectively. $D_v$, $L$, $D_a$ and $J$ denote the number of dimensions of the image features, the number of frames of input image features, the number of dimensions of the audio features, the number of bases of each dictionary, respectively.

The source dictionary and the target dictionary consist of the same utterances, like parallel training data of VC. Input lip images without any voice are converted to image features. These features are represented by a linear combination of bases from the source dictionary and its activities using NMF. Because the source dictionary and the target dictionary are parallel, the estimated activities are multiplied to the target dictionary and the audio features are synthesized.

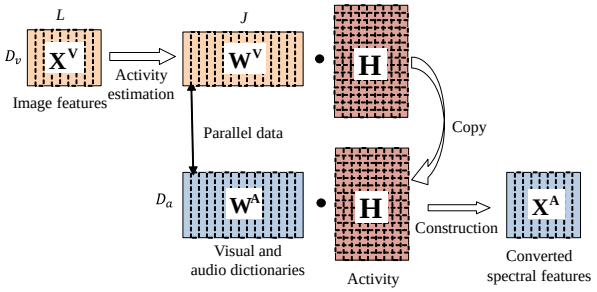


**Fig. 1**. Flow of the proposed method

### 4.2. Dictionary Construction

Fig. 2 shows how to construct the source dictionary and the target dictionary. In this paper, we use images recorded using a high-speed

camera, which made it possible for both the images and the audio to have a high frame rate. For visual features, a two-dimensional Discrete Cosine Transform (DCT) of lip motion images of the source speaker's utterance is used, and a zigzag scan is used to obtain the 1D-DCT coefficient vector. Then, a constant value was added to satisfy the non-negativity constraint of NMF not to change the scale of frame data [19]. In this paper, we employed STRAIGHT [31] for feature extraction and speech synthesis, and use STRAIGHT spectrum for audio features.

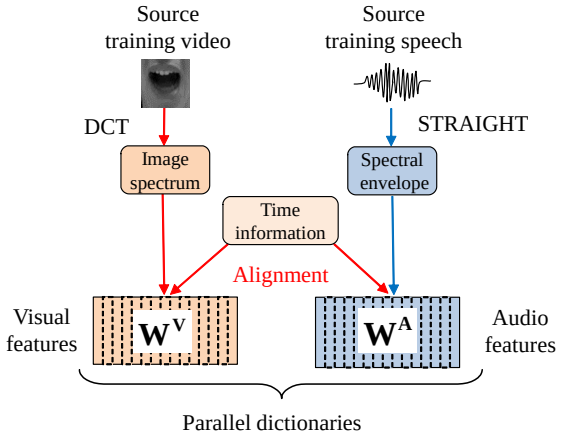


**Fig. 2**. Dictionary construction

### 4.3. Locality Constraint [11]

Locality constraint is adopted to activities in this paper in order to increase the sparsity. The activity is initialized with locality constraint and then update by using (5).

The distance between an input vector and a basis of the source dictionary is defined as follows:

$$\Delta_{j,l} = \sqrt{(\mathbf{x}_l - \mathbf{w}_j)^2} \tag{7}$$

where $\mathbf{x}_l$ and $\mathbf{w}_j$ denote the $l$-th frame of the input vector and the $j$-th basis of the dictionary, respectively. $N$ nearest bases are chosen from all the bases.

$$\mathbf{S}_l = \mathbf{nbest}_{\Delta_l}(\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_J) \tag{8}$$

$$= \mathbf{nbest}_{\Delta_l}(\mathbf{W}) \tag{9}$$

where $\mathbf{S}_l$ denotes a set of bases which consists of $N$ nearest bases to the $l$-th input vector. The activity which relates to $\mathbf{S}_l$ is initialized with a small value and the other activity is initialized with 0. Thus, we can estimate activities which consist of $N$ nearest bases.

## 5. EXPERIMENTAL RESULTS

### 5.1. Experimental Conditions

We recorded 26 utterances of clean continuous Japanese digit speech of one Japanese male by using a high-speed camera. The texts of utterances were taken from CENSREC-1-AV [32] database. Table 1 shows the contents of the database. We used 6 utterances (from a total of 26 utterances) as test data. In closed experiments, 26 utterances including test data were used for the dictionary construction. The number of frames of each dictionary was 30,784. In open experiments, 20 utterances, which do not include test data, were used for the dictionary. The number of frames of each dictionary was 24,368.

**Table 1**. Contents of the database

| number of digits | total number of utterances |
|---|---|
| 2 | 9 |
| 3 | 7 |
| 4 | 10 |
| total | 26 |

Audio and visual data were recorded at the same time in a quiet room. The position of the camera was 65 cm from the speaker and 130 cm from the floor.

The frame rate of the visual data was $1,000$ fps and the image size is $130 \times 80$. Fig. 3 shows examples of lip images recorded by the high-speed camera. For input features, 200-dimensional DCT coefficients of lip motion images of the source speaker's utterance are used. We introduced the segment features for the DCT coefficient, which consist of its consecutive frames (the 2 frames coming before and the 2 frames coming after). Therefore, the total dimension of visual feature is 1,000.

Sampling frequency of target speech was 8kHz and the frame shift was 1ms. Audio spectrum was extracted by STRAIGHT from the training data. The number of dimension of audio spectrum was 513.

We conducted objective and subjective evaluations. In the objective evaluation, Mel-cepstrum Distortion (Mel-CD) between synthesized and target speech is calculated. Mel-CD is calculated by the following equation:

$$\text{Mel-CD[dB]} = \frac{10}{\ln 10}\sqrt{2\sum_{d=1}^{24}(mc_d - \hat{mc}_d)^2} \tag{10}$$

where $mc_d$ and $\hat{mc}_d$ denote the $d$-th dimension of mel-cepstral coefficient of the target and synthesized speech, respectively.

In the subjective evaluation, we conducted a Mean Opinion Score (MOS) test and a dictation test. In an MOS test, subjects evaluated the synthesized voice for speech quality on a 5-point scale (5: excellent, 4: good, 3: fair, 2: poor, 1: bad). In a dictation test, subjects wrote down the converted utterances. These tests were carried out with 7 subjects.

**Fig. 3**. Lip images

### 5.2. Results and Discussions

Fig. 4 shows the mel-cepstral distortions in the evaluation set as a function of the locality. $\beta$ was set to 1 in this evaluation. In the closed experiment, the best result was obtained when the number of bases (locality) was set to 50. As shown in this figure, the distortion increased as the number of bases increased. We assume that this

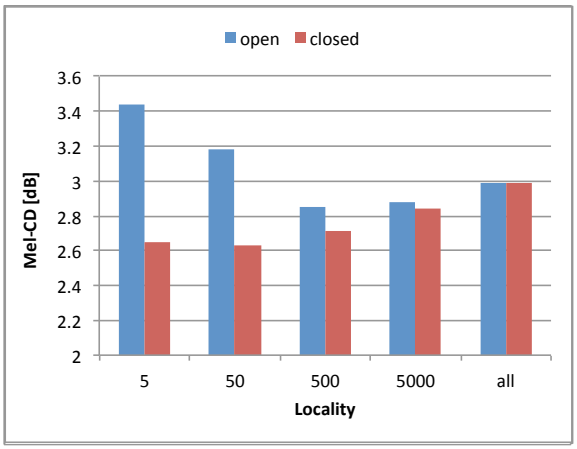


**Fig. 4**. Mel-cepstral distortion as a function of the locality

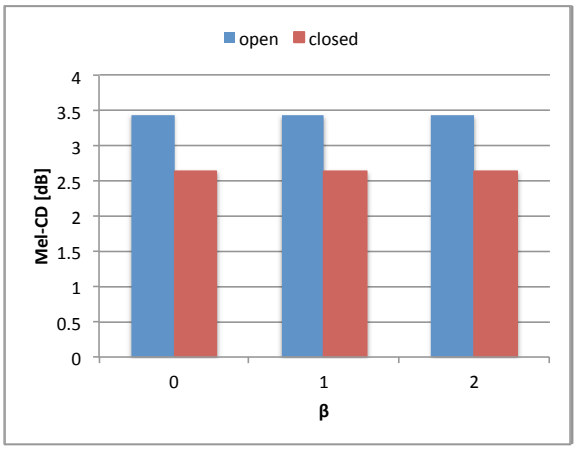


**Fig. 5**. Mel-cepstral distortion as a function of $\beta$

is because an unnecessary basis is included when the locality is increased. In the open experiment, the best result was obtained when the number of bases was set to 500. As shown in this figure, the distortion increased as the number of bases decreased. We assume that this is occurred due to the selection error of local bases.

Fig. 5 shows the mel-cepstral distortions in the evaluation set as a function of $\beta$ in (5). In this evaluation, the locality was set to 5. As shown in this figure, there are no significant differences in these distortions. We assume that this is because of locality-constraint.

Fig. 6 shows the results of the MOS test in the evaluation set as a function of the locality. $\beta$ was set to 1 in this evaluation. In the closed test, the best score was obtained when the locality was 50. In the open test, the best score was obtained when the locality was 5. These results show the effectiveness of locality constraint.

Fig. 7 shows the results of the dictation test. $\beta$ was set to 1 in this evaluation. In the closed experiment, the recognition rate was over 60% when the locality constraint was introduced. In the open experiment, the recognition rate was about 50% when the locality constraint was introduced. These results also show the effectiveness of locality constraint.

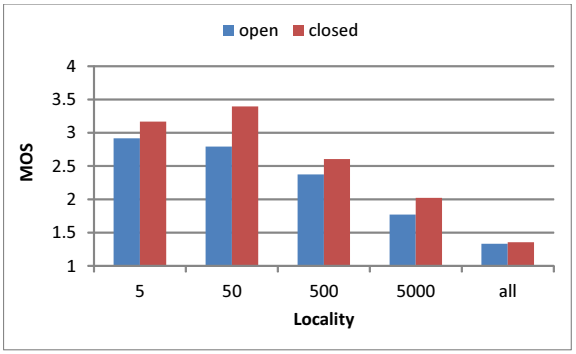


**Fig. 6**. MOS test as a function of the locality

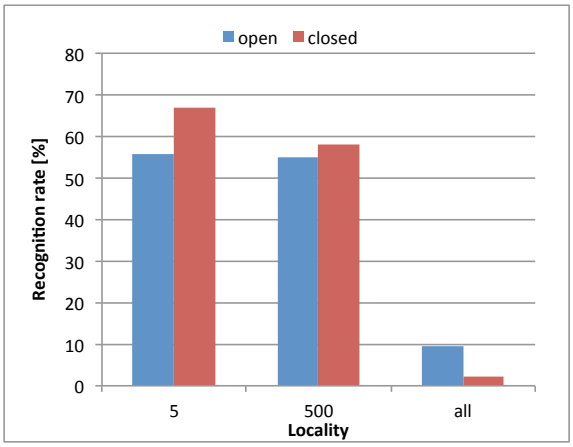


**Fig. 7**. Recognition rate as a function of the locality

## 6. CONCLUSIONS

This paper proposed a lip-to-speech synthesis method that produces speech from lip images without the voice, where lip images and voices are stored as the source dictionary and the target dictionary, respectively. Input images are represented by a linear combination of the basis from the source dictionary. The selected bases are replaced with the corresponding target basis and the speech is synthesized. In this paper, we employed NMF using the $\beta$ divergence and introduced locality-constraint in order to increase the sparseness of the activity matrix. Our objective and subjective evaluations show that our proposed method effectively converted lip images to speech spectra and the effectiveness of locality-constraint was confirmed.

Some problems remain with this method. The proposed method requires high computational times to estimate activities. Virtanen *et al.* [33] proposed an active-set method for NMF that effectively estimates the activity matrix from the over-complete dictionary, and we proposed VC using this method [34]. In future work, we will investigate the optimal number of bases and adopt an active-set method. In this study, there was only one subject person, so in future experiments, we will increase the number of subjects and further examine the effectiveness of our method.

## 7. REFERENCES


[1] C. Veaux, J. Yamagishi, and S. King, "Using HMM-based speech synthesis to reconstruct the voice of individuals with degenerative speech disorders," *in Proc. Interspeech*, pp. 1–4, 2012.

[2] 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.

[3] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Speaking aid system for total laryngectomees using voice conversion of body transmitted artificial speech," *in Proc. Interspeech*, pp. 148–151, 2006.

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

[5] A. Verma, T. Faruquie, C. Neti, S. Basu, and A. Senior, "Late integration in audio-visual continuous speech recognition," *in Proc. ASRU*, 1999.

[6] K. Palecek and J. Chaloupka, "Audio-visual speech recognition in noisy audio environments," *in Proc. International Conference on Telecommunications and Signal Processing*, pp. 484–487, 2013.

[7] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," *Neural Information Processing System*, pp. 556–562, 2001.

[8] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," *in INTERSPEECH*, 2006.

[9] W. Dong, L. Zhang, G. Chi, and X. Wu, "Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization," *IEEE Trans. on Image Processing*, vol. 20, no. 7, pp. 1838–1856, 2011.

[10] R. Takashima, T. Takiguchi, and Y. Ariki, "Exemplar-based voice conversion in noisy environment," *in Proc. SLT*, pp. 313–317, 2012.

[11] R. Aihara, T. Takiguchi, and Y. Ariki, "Individuality-preserving voice conversion for articulation disorders using locality-constrained nmf," *in Proc. Workshop on Speech and Language Processing for Assistive Technologies*, pp. 3–8, 2013.

[12] C. Févotte, N. Bertin, and J. L. Durrieu, "Nonnegative matrix factorization with the itakura-saito divergence. with application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009.

[13] S. Eguchi and Y. Kanno, "Robustifying maximum likelihood estimation," Tech. Rep., Instisute of Statistical Mathematics, 2001.

[14] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the beta-divergence," *Neural Computation*, vol. 23, no. 9, pp. 2421–2456, 2011.

[15] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Efficient algorithms for multichannel extensions of itakura-saito nonnegative matrix factorization," *in Proc. ICASSP*, pp. 261–264, 2012.

[16] J. F. Gemmeke, T. Viratnen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2067–2080, 2011.

[17] R. Aihara, R. Takashima, T. Takiguchi, and Y. Ariki, "Noise-robust voice conversion based on sparse spectral mapping using non-negative matrix factorization," *IEICE Transactions on Information and Systems*, vol. E97-D, no. 6, pp. 1411–1418, 2014.

[18] R. Aihara, R. Takashima, T. Takiguchi, and Y. Ariki, "A preliminary demonstration of exemplar-based voice conversion for articulation disorders using an individuality-preserving dictionary," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2014:5, doi:10.1186/1687-4722-2014-5, 2014.

[19] K. Masaka, R. Aihara, T. Takiguchi, and Y. Ariki, "Multimodal exemplar-based voice conversion using lip features in noisy environments," *in Proc. INTERSPEECH*, vol. 1159-1163, 2014.

[20] A. Barbulescu, T. Hueber, G. Bailly, and R. Ronfard, "Audio-visual speaker conversion using prosody features," *in Proc. AVSP, 12th International Conference on Auditory-Visual Speech Processing*, pp. 11–16, 2013.

[21] K. Sawada, M. Takehara, S. Tamura, and S. Hayamizu, "Audio-visual voice conversion using noise-robust features," *in Proc. ICASSP*, pp. 7899–7903, 2014.

[22] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition," *in Proc. Signals, Systems and Computers*, vol. 1, pp. 40–44, 1998.

[23] I. Ramirez, P. Sprechmann, and G. Sapiro, "Classification and clustering via dictionary learning with structured incoherence and shared features," *in Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3501–3508, 2010.

[24] F. Tao and C. Busso, "Lipreading approach for isolated digits recognition under whisper and neutral speech," *in Proc. Interspeech*, pp. 1154–1158, 2014.

[25] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, "Lipreading using convolutional neural network," *in Proc. Interspeech*, pp. 1149–1153, 2014.

[26] B. Denby, T. Schultz, K. Honda, T. Hueber, and J. M. Gilbert, "Silent speech interfaces," *Speech Communication*, vol. 52, no. 4, pp. 270–287, 2010.

[27] E. Yamamoto, S. Nakamura, and K. Shikano, "Lip movement synthesis from speech based on hidden markov models," *Speech Communication*, vol. 25, no. 1-2, pp. 105–115, 1998.

[28] 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.

[29] X. Zhuang, L. Wang, F. Soong, and M. Hasegawa-Johnson, "A minimum converted trajectory error (mcte) approach to high quality speech-to-lips conversion," *in Proc. INTERSPEECH*, pp. 1736–1739, 2010.

[30] T. Toda, A. Black, and K. Tokuda, "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.

[31] H. Kawahara and H. Matsui, "Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation," *in Proc. ICASSP*, vol. I, pp. 256–259, 2003.

[32] S. Tamura, C. Miyajima, N. Kitaoka, K. Takeda, T. Yamada, T. Takiguchi, S. Tsuge, K. Yamamoto, T. Nishiura, M. Nakayama, Y. Denda, M. Fujimoto, S. Matsuda, T. Ogawa, S. Kuroiwa, and S. Nakamura, "CENSREC-1-AV, an evaluation framework for multimodal speech recognition," Tech. Rep. 7, SLP, 2010.

[33] T. Virtanen, B. Raj, J. F. Gemmeke, and H. Van Hamme, "Active-set newton algorithm for non-negative sparse coding of audio," *in Proc. ICASSP*, pp. 3116–3120, 2014.

[34] R. Aihara, R. Ueda, T. Takiguchi, and Y. Ariki, "Exemplar-based emotional voice conversion using non-negative matrix factorization," *in Proc. APSIPA*, 2014.