Speech Estimation in Non-Stationary Noise Environments Using Timing Structures between Mouth Movements and Sound Signals

Hiroaki Kawashima[1], *Yu Horii*[1], *Takashi Matsuyama*[1]

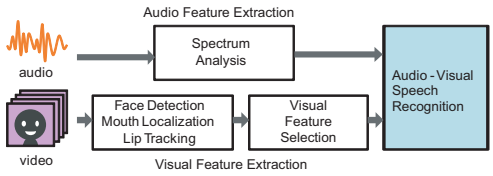
[1]Graduate School of Informatics, Kyoto University, Japan

kawashima@i.kyoto-u.ac.jp, horii@vision.kuee.kyoto-u.ac.jp, tm@i.kyoto-u.ac.jp

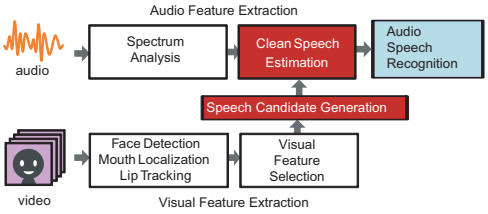
Abstract

A variety of methods for audio-visual integration, which integrate audio and visual information at the level of either features, states, or classifier outputs, have been proposed for the purpose of robust speech recognition. However, these methods do not always fully utilize auditory information when the signal-to-noise ratio becomes low. In this paper, we propose a novel approach to estimate speech signal in noise environments. The key idea behind this approach is to exploit clean speech candidates generated by using *timing structures* between mouth movements and sound signals. We first extract a pair of feature sequences of media signals and segment each sequence into temporal intervals. Then, we construct a cross-media timing-structure model of human speech by learning the temporal relations of overlapping intervals. Based on the learned model, we generate clean speech candidates from the observed mouth movements.

Index Terms: multimodal, non-stationary noise, timing, linear dynamical system, particle filtering



(a) Standard audio-visual speech recognition (AVSR)



(b) Framework using estimated speech from visual cues

Figure 1: Existing and proposed frameworks

1. Introduction

In order to address the increasing demand for robust speech recognition under noise environments, a variety of audio-visual speech recognition (AVSR) methods that utilize visual information have been introduced. These methods first extract features from mouth movements and speech signals. Then, they integrate the features at one of several levels. Early integration, which concatenates multimedia features as the input of a classifier, and late integration, which merges the results from two independent classifiers of visual and audio signals, are standard techniques for the integration [1] (Fig. 1(a)). State-level integration methods, such as coupled hidden Markov models (CHMMs) [2], represent the cooccurrence between two media signals as a dynamic Bayesian framework.

Speech input for AVSR applications is, however, usually affected by a certain level of noise, and because most of AVSR approaches treat both audio and visual media signals on an equal basis, they do not always fully utilize auditory information when the signal-to-noise ratio (SNR) becomes low. For example, CHMMs are often used with the relative reliability of audio and video features, and the weight of audio streams is decreased with the increase of noise ratio. As a result, the recognition performance often relies on only visual information when the noise is nonnegligible.

In this paper, we present a novel method to directly retrieve clean speech features by exploiting visual features. As shown in Fig. 1(b), we first generate multiple speech candidates from captured mouth movements, and then evaluate the consistency of the candidates using simultaneously captured audio signals. Finally, the clean speech signal estimated from using this method can be used as the input for subsequent ASR or AVSR classifiers. This study focuses on the steps for the estimation of clean speech signals in this framework.

In direct estimation methods such as that presented in this paper, it is necessary to have methods for generating precise speech signals from visual features. In particular, we need to carry out the following issues.

1. Reduce the number of generated candidates because one mouth movement usually corresponds to several sounds.
2. Generate continual and smooth speech signals; this is required to check the consistency of the candidate sequence with the input audio signal.
3. Manage the temporal differences between the mouth movements and speech sounds because they are sometimes synchronized loosely.

To satisfy these requirements, we use a hybrid dynamical system (HDS) as a model for each audio and visual signal. HDSs are integrated models of discrete-event systems (e.g., HMMs) and dynamical systems. There are various types of HDSs; however, in this study we use an extended version of the segmental models [3], in which each segment is modeled by a linear dynamical system (LDS). Once we represent the temporal dynamics of each media feature as the switching between multiple LSDs, the temporal relation and cooccurrence of multimedia feature sequences can be extracted as the temporal differences between those switching time points of the constituent LDSs [4]. We build a simple statistical model to represent the temporal relations, which we refer to as the *cross-media timing structure model* [4].

Because HDSs can be trained from input signals in a bottom-up manner, systems can automatically find appropriate sets of motion or sound elements from input media signals, in which each elements is represented by an LDS. Thus, each HDS successfully partitions an input signal (or a feature sequence) into a symbolic sequence with the labels of constituent LDSs; as a consequence, it reduces the calculation cost in the candidate generation step, as required in the previously mentioned requirement 1. Moreover, with regard to requirement 2, HDS can generate a smooth temporal sequence; and with regard to requirement 3, the trained timing structure model can successfully take the systematic temporal gaps and fluctuation between two media signals into account.

1.1. Problem Setting

In this paper, we assume the following specific situation; these can be applicable to, for example, driving environments.

- Use of a single camera and microphones
- No occlusions of the mouth regions in captured images
- Speech affected by non-stationary additive noise
- Availability of users dependent speech data for learning

While these assumptions can be easily extended by using existing techniques of microphone arrays, spectral analysis, and so on, we concentrate on this simple configuration to evaluate the basic properties of our proposed method.

Assuming that a reliable visual feature sequence V is available, we generate candidates of clean speech feature sequences $\hat{S}^{(c)}$ $(c = 1, ..., C)$ from V; then, we evaluate the consistency of the candidate sequences with input audio signal X affected by additive noise N; and finally, we estimate a clean speech feature sequence S. The overall flow of the proposed method consists of three phases: learning, candidate generation, and noise compensation.

2. Candidate generation based on HDSs

Figure 2 shows an overview of the learning and candidate generation phases. In the learning phase, we extract feature sequences of speech signals and mouth movements under a low-noise environment. We train two HDS models from each of the speech and video feature sequences. In the following, we use HDS_s and HDS_v to denote the trained HDSs from speech and visual features, respectively. Then, we train a timing structure model between those features based on the method proposed in [4]. As a result of training of HDSs, the captured multimedia signals are partitioned and represented by pairs of interval sequences. Let $I_k^{(s)} = [b_k^{(s)}, e_k^{(s)}]$ and $I_{k'}^{(v)} = [b_{k'}^{(v)}, e_{k'}^{(v)}]$ be overlapped intervals that appear in the speech and video interval sequences. Let $m_k^{(s)}, m_{k'}^{(v)}$ be labels of LDSs that represent the dynamics in the intervals. Assuming that HDS_s and HDS_v comprise sets of LDSs $\mathcal{D}^{(s)} = \{D_i^{(s)}\}_{i=1}^{N_s}$ and $\mathcal{D}^{(v)} = \{D_{i'}^{(v)}\}_{i'=1}^{N_v}$, respectively, we calculate the following distributions for all the pairs of LDSs $(D_i^{(s)}, D_{i'}^{(v)})$:

$$P(m_k^{(s)} = D_i^{(s)}, m_{k'}^{(v)} = D_{i'}^{(v)} \mid I_k^{(s)} \cap I_{k'}^{(v)} \neq \emptyset), \quad (1)$$

$$P(b_k^{(s)} - b_{k'}^{(v)}, e_k^{(s)} - e_{k'}^{(v)} \mid m_k^{(s)}, m_{k'}^{(v)}, I_k^{(s)} \cap I_{k'}^{(v)} \neq \emptyset) \quad (2)$$

The first distribution represents the cooccurrence of LDSs, and the second distribution represents the possible degree of temporal gaps between two LDSs. We refer to the set of these distributions as the timing structure model and use Φ to represent all the model parameters.

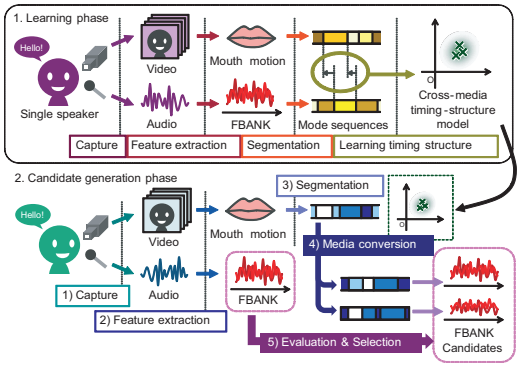


Figure 2: Flow for generating clean speech candidates

In the candidate generation phase, we first extract feature sequences from novel audio and visual input. Let $V = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{T_v}]$ be a captured visual feature sequence. We generate candidates of clean speech sequences $\hat{S}_c = [\hat{\boldsymbol{s}}_1^{(c)}, \ldots, \hat{\boldsymbol{s}}_{T_s}^{(c)}]$ from V. This generation technique is almost similar to the method described in [4] except that the generated sequences are multiple and the modality is inverted; that is, the original method generates a lip motion sequence from a given input audio signal.

The following are the steps for generating a single candidate:

1. Partition the input sequence V into an interval sequence $\mathcal{I}^{(v)} = \{I_1^{(v)}, ..., I_{K_v}^{(v)}\}$ by using the trained HDS_v.
2. Generate an interval sequence $\mathcal{I}^{(s)} = \{I_1^{(s)}, ..., I_{K_s}^{(s)}\}$ from $\mathcal{I}^{(v)}$ based on the trained timing structure model.
3. Generate a speech feature sequence candidate $\hat{S}$ from the generated interval sequence $\mathcal{I}^{(s)}$ by using the trained HDS_s.

Because HDSs are generative models, steps 1 and 3 are performed using standard techniques, such as those described in [4]. Step 2 is the key step in candidate generation. In the original method, the Viterbi algorithm was used to solve the optimization problem:

$$\hat{\mathcal{I}}^{(s)} = \arg\max_{\mathcal{I}^{(s)}} P(\mathcal{I}^{(s)} | \mathcal{I}^{(v)}, \Phi). \quad (3)$$

From the trained timing structure model with parameters Φ, a single interval sequence $\mathcal{I}^{(s)}$ can be estimated from the given $\mathcal{I}^{(v)}$. However, because one mouth movement corresponds to several speech sounds, we utilize the parallel list Viterbi algorithm [5] to find multiple interval sequences. Finally, we generate multiple candidates of speech feature sequences by switching corresponding LDSs in the HDS_s based on each of the generated interval sequences during step 3.

3. Noise compensation using particle filters

In the noise compensation phase, we estimate a clean speech signal by using the generated candidates $\hat{S}_c$ discussed in the previous section together with the observed input audio X. We apply a method that uses a particle filtering technique to trace non-stationary noise signals [6]. Let $\boldsymbol{x}_t, \boldsymbol{s}_t$, and $\boldsymbol{n}_t$ be logarithmic mel-spectrum features of observed audio, clean speech, and noise signals, respectively. Then, this method assumes that

non-stationary noise is modeled by a random walk process:

$$\boldsymbol{n}_{t+1} = \boldsymbol{n}_t + \boldsymbol{\omega}_t. \quad (4)$$

The observation (i.e., audio signal) is acquired by the sum of (unobservable) noise and clean speech signals:

$$\begin{aligned} \boldsymbol{x}_t &= \log(\exp(\boldsymbol{s}_t) + \exp(\boldsymbol{n}_t)) + \boldsymbol{v}_t \\ &= \boldsymbol{s}_t + \log(\mathbf{1} + \exp(\boldsymbol{n}_t - \boldsymbol{s}_t)) + \boldsymbol{v}_t \\ &= f(\boldsymbol{s}_t, \boldsymbol{n}_t) + \boldsymbol{v}_t, \end{aligned} \quad (5)$$

where $\boldsymbol{v}_t$ is a model error component (e.g., reverberation) with Gaussian distribution $\mathcal{N}(\mathbf{0}, \Sigma_x)$, Σ_s is a covariance matrix of $\boldsymbol{s}_t$, $\mathbf{1}$ is a vector filled by 1, and $f(\boldsymbol{s}_t, \boldsymbol{n}_t)$ is defined as

$$f(\boldsymbol{s}_t, \boldsymbol{n}_t) \triangleq \boldsymbol{s}_t + \log(\mathbf{1} + \exp(\boldsymbol{n}_t - \boldsymbol{s}_t)).$$

The equations (4) and (5) constitute a non-linear state-space model, and the noise can be inferred by particle filtering techniques [6, 7]. Finally, we can estimate clean speech based on a minimum mean square error (MMSE)-based method [8]:

$$\hat{\boldsymbol{s}}_t = \boldsymbol{x}_t - \sum_{l=1}^{L_s} P(l|\boldsymbol{x}_t)(f(\boldsymbol{\mu}_{s,l}, \boldsymbol{n}_t) - \boldsymbol{\mu}_{s,l}),$$

$$P(l|\boldsymbol{x}_t) = \frac{w_{s,l}\mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_{x,l}, \Sigma_{x,l})}{\sum_{m=1}^{L_s} w_{s,m}\mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_{x,m}, \Sigma_{x,m})},$$

where $\boldsymbol{\mu}_{x,l}$ is the mean vector and $\Sigma_{x,l}$ is the covariance matrix of observation model (5), which can be approximately calculated by using the vector Taylor series method [9].

The nonlinear observation model shown in equation (5) has a clean speech $\boldsymbol{s}_t$ as the input component. Because the input needs to be determined on the basis of prior knowledge of clean speech, this component is sampled from the following Gaussian mixture model (GMM) in the original method [6]:

$$p(\boldsymbol{s}_t) = \sum_{l=1}^{L_s} w_{s,l}\mathcal{N}(\boldsymbol{\mu}_{s,l}, \Sigma_{s,l}), \quad (6)$$

where L_s is the number of Gaussians and $w_{s,l}$ are the weights for each of the constituent Gaussians.

However, when this type of a static GMM is used, the noise tracking often fails under low SNR environment when the degree of noise is large. The key principle in our proposed method is to use the dynamically changing GMM, shown in the following equation, by exploiting the speech candidates $\hat{S}_c$ generated from the input visual feature sequence V.

$$p(\boldsymbol{s}_t|V) = \sum_{c=1}^{C} W_c \mathcal{N}(\hat{\boldsymbol{s}}_t^{(c)}, \Sigma_{s_c}), \quad (7)$$

where the weight W_c can be determined by using the likelihood calculated in the parallel list Viterbi algorithm during the optimization in equation (3).

4. Experiments

We evaluated the basic capabilities of our proposed method by using speech data obtained from a male subject. Five speech sequences, each of which comprises isolated sounds of the five vowels /a//i//u//e//o/, were captured together with the speaker's facial images. The resolution and frame rate of the images were 640 × 480 and 60 fps, respectively. Figure 3(a) shows an example of a captured image. The quantization size and sampling rate of audio data were 16 bit and 48 kHz, respectively, which was downsampled to 16 kHz afterwards.

(a) Captured image (b) Tracked feature points (c) Mouth region image

Figure 3: (a) Example of a captured image, (b) extracted feature points by AAM, (c) extracted mouth region.

4.1. Visual and speech feature extraction

Visual features are required to describe the difference between phonemes as much as possible, and thus, the shapes around the lips are inadequate for providing such visual information. Therefore, we use appearance-based features around the mouth. We tracked facial feature points by using the active appearance model (AAM) [10] (Fig 3(b)), cropped the mouth region by estimating the center of the lip shapes, and normalized the size of the mouth region. Then the region was downsampled to a resolution of 32 × 32, and the peripherals of each image were smoothly masked (Fig 3(c)). Principal component analysis was applied, and the top 20 principal components were used in a visual feature vector.

We used the HMM Tool Kit (HTK) to extract filter bank coefficients (FBANK) as speech features for speech candidates and during the noise compensation algorithm. We also extracted line spectrum pairs (LSPs) [11] as supplemental features during the training steps of HDS_a because LSPs have shown to be advantageous for representing smooth temporal sequences. The window size and step size used in the short-term spectrum analysis was about 25 ms and 16.7 ms, respectively, which was synchronized with the frame rate of visual features to simplify the implementation. The length of each sequence was 300 frames.

4.2. Generation of speech candidates

Two HDSs were trained from the extracted sequences via the learning algorithm proposed in [12], and the numbers of LDSs in HDS_v and HDS_s were manually determined to be 8 and 10, respectively, by examining the modeling error curves. The used LDSs were second-order vector autoregressive models.

A timing structure model was trained from the pairs of interval sequences obtained in the training step of HDSs. First, every overlapping interval pair was extracted to calculate a probability table in equation (1). Then, the samples of temporal differences between their beginning points and between the end points (i.e., $(b_k^{(s)} - b_{k'}^{(v)}, e_k^{(s)} - e_{k'}^{(v)})$) were extracted, convolved with a Gaussian kernel (standard deviation was 3 frames), and accumulated onto two dimensional spaces to obtain temporal difference distributions in equation (2).

From a visual feature sequence consisting of the five vowels, we generated 50 speech sequences candidates by using the method described in section 2, and ranked the candidates on the basis of the likelihood obtained from the parallel list Viterbi algorithm. Figure 4 shows the 1st and the 26 th candidates. We see that the 1st candidate is more similar to the original clean speech (top) than the 26 th candidate.

4.3. Noise compensation using average candidates

We performed the noise compensation (clean speech estimation) method described in section 3 based on the leave-one-out cross validation on the five captured sequences used in the previous subsection. For each of the five tests, a pair of

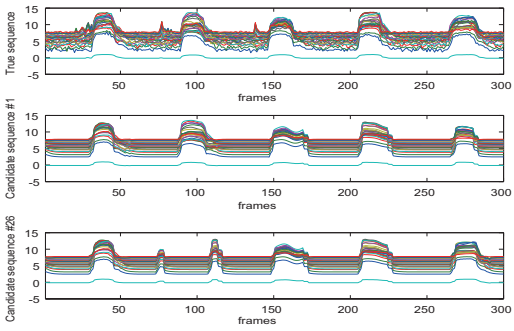


Figure 4: Example of a generated FBANK sequence from a visual feature vector. (Top) clean speech (ground truth), (middle) generated candidate whose likelihood is the highest, (bottom) generated candidate whose likelihood is 26 th.

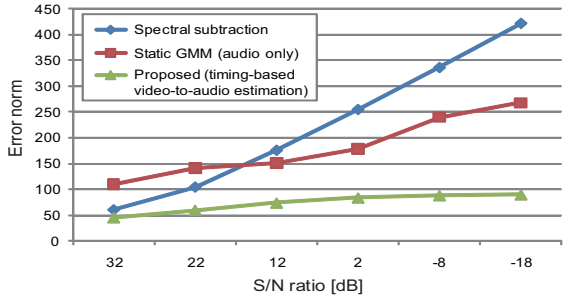


Figure 5: Error norms between original and estimated speech feature sequences via six levels of SNRs.

HDSs was trained from four pairs of clean speech and visual feature sequences. Then, the remaining pair of feature sequences were used as test data. The audio test sequences were prepared by adding non-stationary noise obtained from JEITA noise database [13]. The added noise captured in a factory included the sound of several industrial machines such as an air wrench, and had strong non-stationarity. We prepared speech with six different SNRs: $-18, -8, 2, 12, 22$, and 32 dB; and we then extracted speech features as the observation sequences $X = \{x_t\}$ of equation (5). The GMM parameters were trained by HTK from clean speech data. The number of Gaussians used in the speech GMM was 13.

We compared our proposed method with two existing methods: spectral subtraction (specsub in VOICEBOX [14]) and particle filtering with static GMM in equation (6). To examine the basic properties of the proposed method, we used an average of 50 candidates and constructed a single time-varying Gaussian distribution for the model shown in equation (7). The number of particles used was 50, and the covariance matrices are $\Sigma_\omega = \mathrm{diag}(0.01)$ and $\Sigma_{s_c} = \mathrm{diag}(1.0)$. Figure 5 shows the error norms between the original clean speech and the estimated speech signals. To reduce the effect of random sampling, we iterated the particle filtering process three times. Therefore, each error norm of particle filtering methods in the figure was calculated as the average of 15 error norms (5 test data $\times$ 3 iteration). We see that the estimated speech feature sequences are much closer to the original sequences compared to existing method, even when the SNR is closer to 0 dB.

5. Conclusions

In this paper, we have proposed a novel speech estimation method that directly estimates feature sequences of clean speech from observed mouth movements. Candidate generation is realized by using hybrid dynamical systems (HDSs) and a timing structure model between the HDSs. While the evaluation presented in this paper is limited, it shows that the proposed method can estimate clean speech with much higher precision under non-stationary noise environments as compared to the methods that use only audio data. Our future work will involve a comparative evaluation of our method with existing AVSR methods by using a large speech corpus as well as a detailed analysis of the characteristics of the proposed method.

6. Acknowledgements

This work is partially supported by the Grant-in-Aid for Scientific Research of the Ministry of Education, Culture, Sports, Science and Technology of Japan under the contracts of 18049046 and 21680016.

7. References

[1] T. Chen and R. R. Rao, "Audio-visual integration in multimodal communication," *Proceedings of the IEEE*, pp. 837–852, 1998.

[2] A. V. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy, "Dynamic Bayesian networks for audio-visual speech recognition," *EURASIP Journal on Applied Signal Processing*, vol. 2002, no. 11, pp. 1–15, 2002.

[3] M. Ostendorf, V. Digalakis, and O. A. Kimball, "From HMMs to segment models: a unified view of stochastic modeling for speech recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 4, no. 5, pp. 360–378, 1996.

[4] H. Kawashima and T. Matsuyama, "Interval-based linear hybrid dynamical system for modeling cross-media timing structures in multimedia signals," *Proc. International Conference on Image Analysis and Processing*, pp. 789–794, 2007.

[5] N. Seshadri and C.-E. Sundberg, "List Viterbi decoding algorithms with applications," *IEEE Transactions on Communications*, 1994.

[6] M. Fujimoto and S. Nakamura, "A non-stationary noise suppression method based on particle filtering and polyak averaging," *IEICE Transactions on Information and Systems*, vol. 89, no. 3, pp. 922–930, 2006.

[7] S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for on-line non-linear/non-gaussian Bayesian tracking," *IEEE Transactions on Signal Processing*.

[8] J. C. Segura, A. D. L. Torre, M. C. Benitez, and A. M. Peinado, "Model-based compensation of the additive noise for continuous speech recognition. experiments using the AURORA II database and tasks," *Proc. EuroSpeech*, vol. 1, pp. 221–224, 2001.

[9] P. Moreno, B. Raj, and R. Stern, "A vector Taylor series approach for environment-independent speech recognition," *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1996.

[10] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance model," *Proc. European Conference on Computer Vision*, pp. 484–498, 1998.

[11] K. Tokuda and SPTK Working Group, *Reference Manual for Speech Signal Processing Toolkit Ver. 3.2*.

[12] H. Kawashima and T. Matsuyama, "Multiphase learning for an interval-based hybrid dynamical system," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. 88, no. 11, pp. 3022–3035, 2005.

[13] S. Itahashi, "A noise database and Japanese common speech data corpus," *The Journal of the Acoustical Society of Japan*, vol. 47, no. 12, pp. 951–953, 1991.

[14] M. Brookes, *VOICEBOX: Speech Processing Toolbox for MATLAB*. [Online]. Available: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html