AUDIO-VISUAL FEATURE INTEGRATION BASED ON PIECEWISE LINEAR TRANSFORMATION FOR NOISE ROBUST AUTOMATIC SPEECH RECOGNITION

Yosuke Kashiwagi[1], Masayuki Suzuki[2], Nobuaki Minematsu[2], Keikichi Hirose[1]

[1]Graduate School of Information Science and Technology,
[2]Graduate School of Engineering,
The University of Tokyo, Japan
{kashiwagi, suzuki, mine, hirose}@gavo.t.u-tokyo.ac.jp

ABSTRACT

Multimodal speech recognition is a promising approach to realize noise robust automatic speech recognition (ASR), and is currently gathering the attention of many researchers. Multimodal ASR utilizes not only audio features, which are sensitive to background noises, but also non-audio features such as lip shapes to achieve noise robustness. Although various methods have been proposed to integrate audio-visual features, there are still continuing discussions on how the vest integration of audio and visual features is realized. Weights of audio and visual features should be decided according to the noise features and levels: in general, larger weights to visual features when the noise level is low and vice versa, but how it can be controlled? In this paper, we propose a method based on piecewise linear transformation in feature integration. In contrast to other feature integration methods, our proposed method can appropriately change the weight depending on a state of an observed noisy feature, which has information both on uttered phonemes and environmental noise. Experiments on noisy speech recognition are conducted following to CENSREC-1-AV, and word error reduction rate around 24% is realized in average as compared to a decision fusion method.

Index Terms— Feature enhancement, Multimodal ASR, SPLICE, noise robustness

1. INTRODUCTION

Although Automatic Speech Recognition achieves high performance in clean environments, recognition rates sharply degrade in low SNR conditions. To deal with this problem, a variety of methods have been proposed. Among them, conversion of noisy speech features to clean speech features, (henceforth, feature enhancement) attains researchers' concern, since several effective techniques have been developed recently, such as Stereo-based Piecewise Linear Compensation for Environments (SPLICE) [1], Vector Taylor Series(VTS) based compensation [2].

When supplemental information such as lip movements, EMG signals, is available during utterances, it can be used to facilitate speech recognition. For instance, voice activity detection from visual information can largely improve the recognition performances [3, 4]. These multimodal ASR approaches can be broadly categorized into two types: decision fusion [5–8] and feature fusion [8, 9] methods. Decision fusion methods combine outputs from single-modality HMM classifiers (audio and visual ones) to recognize speech with visual information. Here, a weighted sum of class conditional log-likelihoods from the two classifiers is used. On the other hand, feature fusion methods combine audio and visual features in feature domain. And HMM classifiers of the combined feature are used to recognize speech. In this paper, we focus on the feature fusion.

An important issue of feature fusion methods is weighting of audio and visual features. We should appropriately decide the weights of these features using various kinds of information such as the type of environmental noises and phonemes. To this end, we propose a new feature fusion method based on piecewise linear transformation. We divide the noisy audio feature space into many states using Gaussian Mixture Model (GMM), and estimate an optimal linear transformation for joint audio-visual features for each state so that the audio-visual features are appropriately weighted depending on the states. This method is inspired by the well-known SPLICE method [1], which utilizes GMM of noisy audio feature and piecewise linear transformation. A difference from SPLICE is utilizing visual information for piecewise linear transformation.

In the following sections, multimodal speech recognition methods are first explained in section 2, followed by a detailed description of our proposal in section 3. Section 4 shows experimental results, and section 5 concludes the paper.

2. MULTIMODAL SPEECH RECOGNITION

Due to distortions of speech features by noise, the performance of ASR often degrades in low SNR conditions. Supplemental features other than audio ones can solve this situation if these features are less influenced by the noise. This type of ASR is called multimodal ASR, which often uses visual features obtainable during utterances, such as lip shapes, facial expressions, and so on.

When using visual features, the main question is how audio and visual features should be incorporated in the recognition process. As mentioned in Section 1, audio and visual information can be integrated in two ways for ASR: feature fusion and decision fusion.

2.1. Feature fusion

Feature fusion methods generate concatenated vectors from audio feature vectors and visual feature vectors, and train a "multi-modal" HMM classifier. The feature fusion is advantageous in that it can utilize the correlation between multiple features from different modalities which helps in better task accomplishment. Also, it requires only one learning phase on the combined feature vector. However, increasing feature dimension influence the performance of the feature fusion. To reduce dimensionality of concatenated vector, principle component analysis (PCA) or linear discriminant analysis (LDA) is often applied.

2.2. Decision fusion

In contrast, decision fusion methods combine log-likelihoods from audio and vision classifiers. The methods independently model audio-only and visual-only expressions, then combines these recognition results at the end. When two audio and visual HMM classifiers are used, the following multi-stream HMM framework can be adopted:

$$\begin{aligned} &L_{av,t} = \lambda_a L_{a,t} + \lambda_v L_{v,t} \\ &\text{where } \lambda_a + \lambda_v = 1, \quad \lambda_a, \lambda_v \geq 0. \end{aligned} \tag{1}$$

$L_{a,t}$ and $L_{v,t}$ are log-likelihoods of audio and visual HMM classifiers for time t, respectively, and $L_{av,t}$ is the integrated one. Since the two outputs are used directly for fusion, good temporal alignment between the two feature streams (audio and visual) is not guaranteed. To solve this problem, the forced alignment on the audio models is often used to train visual HMMs.

3. PROPOSED METHOD

In the feature fusion method, PCA or LDA are often used to reduce the dimensionality and to weight the audio and visual features. However, its weighting may not appropriate because all input data are transformed using a common matrix. The weight should be changed depending on noise environments, phoneme contents, and so on.

To solve the problem, we introduce a new feature fusion method, which combines audio and image information appropriately depending on observed audio features. We estimate audio features of clean speech from features of observed audio and visual ones. In this framework, we can interpret enhanced audio features as the joint audio-visual features after dimension reduction.

We propose a method to estimate a clean feature $\hat{\boldsymbol{x}}$ from a noisy feature $\boldsymbol{y}$ and a visual feature $\boldsymbol{i}$ using a piecewise linear transformation as:

$$\hat{\boldsymbol{x}} = \sum_k P(k|\boldsymbol{y}) \boldsymbol{A}_k \boldsymbol{m}', \tag{2}$$

where $\boldsymbol{m}' = [1, \boldsymbol{y}^\top, \boldsymbol{i}^\top]^\top$ is an augmented joint vector of audio and visual features. As for visual features $\boldsymbol{i}$, appearance features with eigen-face [11] are adopted. Assuming that $\boldsymbol{y}$ follows a Gaussian mixture of the noisy speech cepstra, $P(k|\boldsymbol{y})$ can be represented in the following way:

$$\begin{aligned} P(\boldsymbol{y}|k) &= \mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \\ P(\boldsymbol{y}) &= \sum_k \pi_k \mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \\ P(k|\boldsymbol{y}) &= \frac{\pi_k \mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_k \pi_k \mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}, \end{aligned} \tag{3}$$

where π_k, $\boldsymbol{\mu}_k$, and $\boldsymbol{\Sigma}_k$ are weight, mean, and covariance matrix of k-th element of the GMM, respectively. We train these parameters using training data in advance. Transformation matrices $\{\boldsymbol{A}_k\}$ are obtained using the following weighted minimum mean square error criterion:

$$\{\boldsymbol{A}_k\} = \underset{\{\bar{\boldsymbol{A}}_k\}}{\operatorname{argmin}} \sum_l \sum_k P(k|\boldsymbol{y}_l) \|\boldsymbol{x}_l - \bar{\boldsymbol{A}}_k \boldsymbol{m}'\|^2, \tag{4}$$

where $\boldsymbol{x}_l$ and $\boldsymbol{y}_l$ are time-synchronized features, and l is the index of training data. This equation can be solved analytically as follows:

$$\hat{\boldsymbol{A}}_k = \boldsymbol{X} \boldsymbol{R}_k \boldsymbol{M}'^\top (\boldsymbol{M}' \boldsymbol{R}_k \boldsymbol{M}'^\top)^{-1}, \tag{5}$$

where $\boldsymbol{R}_k$ is a diagonal matirx which has diagonal components $[P(k|\boldsymbol{y}_1), P(k|\boldsymbol{y}_2)], \ldots, P(k|\boldsymbol{y}_L)]]$. $\boldsymbol{M}'$ represents aligned $\boldsymbol{m}'$ vectors.

3.1. Comparison with SPLICE

SPLICE is a speech enhancement method, which estimates features of clean speech $\boldsymbol{x}$ from those of noisy speech $\boldsymbol{y}$ with piecewise linear transformations as follows:

$$\hat{\boldsymbol{x}} = \sum_k P(k|\boldsymbol{y}) \boldsymbol{A}_k \begin{bmatrix} 1 \\ \boldsymbol{y} \end{bmatrix}, \tag{6}$$

where $\hat{\boldsymbol{x}}$ is the estimated feature, and a probability density function of $\boldsymbol{y}$ is assumed to be a GMM. Therefore, our proposed method is very similar to SPLICE. The difference between these two methods is the input of linear transformation part. While SPLICE only uses audio features, our proposed method uses the joint features of audio and visual ones.

4. EXPERIMENTS

Experiments are conducted for the CENSREC-1-AV [12] task, where Japanese digits are recognized in additive noise conditions. Clean speech data are selected from CENSREC-1-AV, and noisy speech data are created by adding noises included in noisex92 [13] in 5 SNR levels from 20 dB to 0 dB. Tables 1 and 2 summarize audio and visual data used for the experiments. Noisy visual data are not used. Table 3 shows audio and visual features used for the experiments. Audio features are 13 dimensional mel-frequency cepstrum coefficients and their Δ and Δ^2 parameters, totally 39 dimensions. Visual features are obtained by conducing PCA on pixel color data obtained by raster scan: 10 dimensional for each color resulting in 30 dimensional for RGB. We linearly interpolated the

Table 1. Audio data.

Samplling rate	16kHz
Auantization bits	16 bit/sample
Audio noise	noisyx-92 5 types (babble,factory1, factory2,car (Volvo),white) 5 SNR levels (20 dB to 0dB)

Table 2. Visual data.

Frame rate	29.97Hz
Pixsel	24 bit color
Data size	81 pixel width $\times$ 55 pixel height
Visual noise	none (clean)

Table 3. Audio and visual features.

Audio	MFCC+Δ+Δ^2
Visual	Raster scan + PCA (10 $\times$ 3 (RGB) = 30 dimensions)

Table 4. Data sets for training and testing.

	clean HMM	multi HMM	Trans.	Test
Speaker	male 22 female 20			male 25 female 26
Audio data	clean	clean noisyx-92 (babble, factory1, factory2, car (Volvo), white) 20dB 15dB 10dB 5dB 0dB		
visual data	clean color (RGB)			

visual features so that visual features are frame-synchronized with audio features.

Word HMMs for speech recognition are trained in two cases: using only clean speech data (henceforth, clean HMM) and using clean speech data and noisy speech data with 5 types of noises in 5 SNR levels (henceforth, multi-condition HMM) together as indicated in Table 4. Clean and noisy speech data in Table 4 are also used for training linear transformation matrices. Testing is conducted using utterances by different speakers. The proposed method is compared with three baseline methods: no enhancement, conventional feature fusion with dimension reduction from 69 to 39 by PCA, and conventional decision fusion which uses the best integration weight among 0.0,0.2,0.4,0.6,0.8,1.0. In decision fusion, good temporal alignment between audio and visual streams is not guaranteed. To reduce the effect of misalignment, visual HMMs are forced-aligned to audio HMMs during training. In particular, the audio and visual model parameters at each state are used for audio-visual HMMs without any changes and the transition matrices of audio HMMs are also used without any changes. Comparison is conducted also with conventional SPLICE to show the effects of using visual information. All the recognition experiments are conducted with CMN (Cepstral Mean Normalization).

Figure 1 compares the word error rates averaged over all the noise types and levels, when clean condition HMMs and multi-condition HMMs are used. For all the cases, recognitions are conducted using HMMs with 39 dimensional feature vectors. It is clear from the figure that, among the baseline methods (no enhancement, feature fusion, decision fusion), the decision fusion performs the best. Our proposed method surpasses the best method (conventional decision fusion), achieving error reduction rates of 25% in clean HMMs and 24% in multi-condition HMMs. Compared with SPLICE, our proposed method also achieves improvements in word error rates. This shows the effectiveness of using visual features for feature enhancement based on a piecewise linear transformation. Figures 2–6 shows word error rates (averaged over all noise levels) in each noise type. In almost all noise types, our proposed method realized the lowest word error rates. Only one exception is car noise, where the decision fusion performs better than our method. This is because the word error rate is low enouth even for non-enhancement case. Although word error rates are not shown for each noise level, the results show similar tendencies.

5. CONCLUSION

In this paper, a multimodal ASR method is proposed based on piecewise linear transformation. We use noisy audio-visual joint features and clean audio features for training the transformation. In the CENSREC-1-AV task, we achieved 25% and 24% error reduction rates as compared with conventional decision fusion method, which performs the best among three baseline methods. Moreover, our proposed method performs better than SPLICE which only uses audio features for piecewise linear transformation. Effectiveness of using visual features for feature enhancement is proved through the experiments.

6. REFERENCES

[1] J. Droppo, *et.al.*, "Evaluation of the SPLICE on the Aurora2 and 3 Tasks," *Proc. ICSLP*, pp. 29–32, 2002.

[2] V. Stouten, "Robust Automatic Speech Recognition in Time-varying Environments," PhD thesis, 2006.

[3] S. Tamura, *et.al.*, "A Robust Audio-Visual Speech Recognition Using Audio-Visual Voice Activity Detection," *Proc. Interspeech*, 2010.

[4] I. Almajai and B. Milner, "Using audio-visual features for robust voice activity detection in clean and noisy speech," *Proc. EUSIPCO*, 2008.

[5] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Trans. Multimedia*, vol. 2, pp. 141–151, 2000.

[6] G. Potamianos and H. P. Graf, "Discriminative training of HMM stream exponents for audio-visual speech recognition," *Proc. ICASSP*, pp. 3733–3736, 1998.

[7] M. Tariquzzaman, *et. al.*, "Performance Improvement of Audio-Visual Speech Recognition with Optimal Reliability Fusio," *ICICIS*, pp. 203–206, 2011

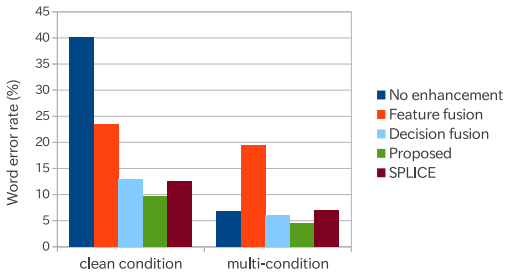

Fig. 1. Word error rates for clean HMMs and multi-condition HMMs.
The results are averaged over all noise types and levels.

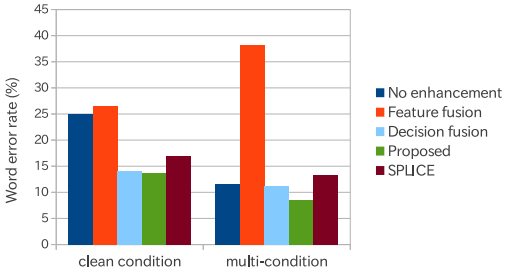

Fig. 2. Word error rates for clean HMMs and multi-condition HMMs in babble noise condition.
The results are averaged over all noise levels.

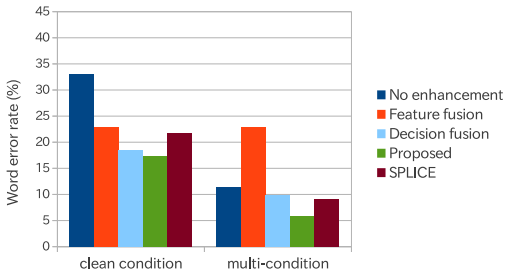

Fig. 3. Word error rates for clean HMMs and multi-condition HMMs in factory1 noise.
The results are averaged over all noise levels.

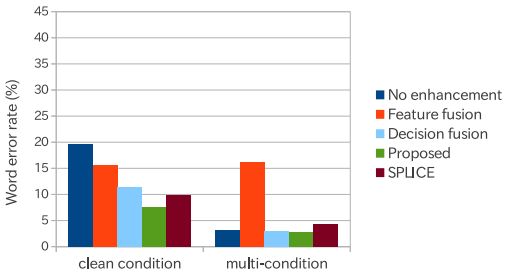

Fig. 4. Word error rates for clean HMMs and multi-condition HMMs in factory2 noise.
The results are averaged over all noise levels.

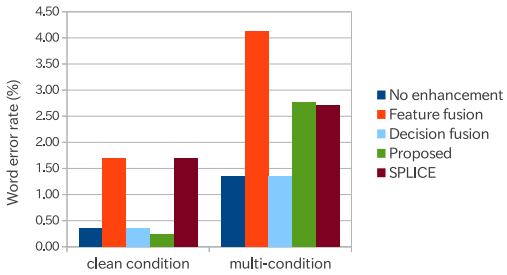

Fig. 5. Word error rates for clean HMMs and multi-condition HMMs in car noise (Volvo noise).
The results are averaged over all noise levels.

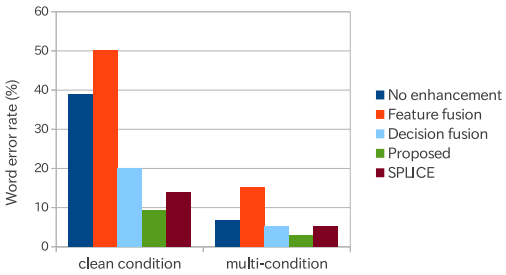

Fig. 6. Word error rates for clean HMMs and multi-condition HMMs in white noise.
The results are averaged over all noise levels.

[8] G. Potamianos, *et. al.*, "Recent advances in the automatic recognition of audiovisual speech," *Proc. IEEE*, pp. 1306–1326, 2003.

[9] G. Potamianos, *et. al.*, "Hierarchical discriminant features for audio-visual LVCSR," *Proc. ICASSP*, pp. 165–168, 2001.

[10] Cootes, T.F., "Active Appearance Model," *Proc. European Conference on Computer Vision,* vol. 2, pp. 484–498, 1998.

[11] Turk, M.A., *et. al.*, "Face recognition using eigenfaces," CVPR, pp. 586–591, 1991.

[12] S. Tamura, *et al.*, "CENSREC-1-AV: An audio-visual corpus for noisy bimodal speech recognition," *Proc. AVSP*, 2010.

[13] http://www.milab.is.tsukuba.ac.jp/corpus/noisedb.html