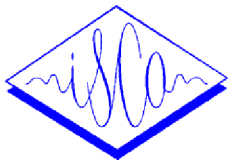
ISCA Archive
http://www.isca-speech.org/archive



AVSP 2010 - International Conference
on Audio-Visual Speech Processing
Hakone, Kanagawa, Japan
September 30-October 3, 2010

CENSREC-1-AV: An audio-visual corpus for noisy bimodal speech recognition

Satoshi TAMURA[1], *Chiyomi MIYAJIMA*[2], *Norihide KITAOKA*[2], *Takeshi YAMADA*[3], *Satoru TSUGE*[4], *Tetsuya TAKIGUCHI*[5], *Kazumasa YAMAMOTO*[6], *Takanobu NISHIURA*[7], *Masato NAKAYAMA*[8], *Yuki DENDA*[9], *Masakiyo FUJIMOTO*[10], *Shigeki MATSUDA*[11], *Tetsuji OGAWA*[12], *Shingo KUROIWA*[13], *Kazuya TAKEDA*[2], *Satoshi NAKAMURA*[11]

[1] Gifu University [2] Nagoya University [3] University of Tsukuba [4] Daido University [5] Kobe University
[6] Toyohashi University of Technology [7] Ritsumeikan University [8] Kinki University [9] Murata Machinery
[10] NTT CS Laboratory [11] NiCT/ATR [12] Waseda University [13] Chiba University

E-mail: tamura@info.gifu-u.ac.jp

Abstract

In this paper, an audio-visual speech corpus CENSREC-1-AV for noisy speech recognition is introduced. CENSREC-1-AV consists of an audio-visual database and a baseline system of bimodal speech recognition which uses audio and visual information. In the database, there are 3,234 and 1,963 utterances made by 42 and 51 speakers as a training and a test sets respectively. Each utterance consists of a speech signal as well as color and infrared pictures around a speaker's mouth. A baseline system is built so that a user can evaluate a proposed bimodal speech recognizer. In the baseline system, multi-stream HMMs are obtained using training data. A preliminary experiment was conducted to evaluate the baseline using acoustically noisy testing data. The results show that roughly a 35% relative error reduction was achieved in low SNR conditions compared with an audio-only ASR method.

Index Terms: audio-visual database, bimodal speech recognition, noise robustness, eigenface, optical flow.

1. Introduction

Automatic Speech Recognition (ASR) has been developed and investigated in order to realize a smart input interface for mobile devices, e.g. laptops, cell phones and personal digital assistants. For example, hands-free speech User Interface (UI) is strongly required in in-car conditions to operate a cell phone or an automotive navigation system. Smartphone becomes recently much popular, and a speech recognition application is installed on the phone to retrieve information. It is well known that ASR has a possibility to innovate UI technology, however, ASR has suffered from the degradation of its performance in noisy or real environments.

Using visual information, that is not affected by acoustic noises, is one of the prospective methods in order to overcome the noise distortion. Bimodal ASR (or audio-visual / multimodal ASR) has attracted attention as a method to ensure the robustness of ASR, and has been researched particularly for this two decades; Potamianos et al. developed a bimodal ASR system using Linear Discriminant Analysis (LDA) and Maximum Likelihood Linear Transform (MLLT) to obtain acoustic and visual features [1]. Since visual feature is crucial for bimodal ASR and lipreading, a number of parameters have been proposed: optical-flow, eigenface, 2DDCT and AAM-based features [2, 3, 4]. Investigation of asynchrony between audio and visual modalities [5], stream weight optimization for multi-stream HMMs which are widely used in this literature [6], and development of real-time bimodal ASR [7] are also major topics.

In order to develop a bimodal ASR method, audio-visual speech databases are widely used where a number of speakers uttered digits, words, and sentences [8, 9]. On the other hand, it is essential to evaluate a recognizer using a common database and a baseline method: an evaluation framework including audio-visual training and testing speech data as well as a baseline system and its results.

A working group of the Corpora and Environments for Noisy Speech RECognition (CENSREC) has been established in the Information Processing Society of Japan (IPSJ) aiming at robust speech recognition in noisy environments. In this paper, we describe our recent efforts to build a new evaluation framework CENSREC-1-AV. This corpus is available for robust bimodal ASR in acoustically and visually noisy environments. CENSREC-1-AV is designed according to the first corpus named CENSREC-1 (AURORA-2J) [10]. More that 5,000 digit utterances were collected by using a lapel microphone and color / infrared cameras. Basic training and testing scripts are also available to compare a user's proposed bimodal ASR method with a baseline system.

The rest of this paper is organized as follows: In Section 2, an audio-visual database in CENSREC-1-AV is introduced. A baseline system is mentioned in Section 3, followed its results in Section 4. Finally Section 5 concludes this paper.

2. Database

2.1. Recording condition

Japanese connected digit utterances were collected. Each utterance consists of 1-7 digits, and each digit is pronounced as *ichi*(1), *ni*(2), *san*(3), *yon*(4), *go*(5), *roku*(6), *nana*(7), *hachi*(8), *kyu*(9), *zero* or *maru*(0), respectively. In an office environment (see Figure 1), a subject sat on a blue background. Speech signals were recorded on a lapel microphone, which is connected to two cameras. One of the cameras captured color movies, and simultaneously, the another captured infrared pictures by using a lens filter. Infrared pictures may be useful when light condition is drastically changed. The sampling frequencies of speech and movie data were 48kHz and 29.97Hz (NTSC) respectively. The image size of the interlaced color and infrared movies was 720×480.

2.2. Training data

A training data set is used to build audio and visual models as well as to compute eigenvectors for visual parameterization. The data set consists of 3,234 utterances in total, spoken by 20 female and 22 male speakers. Clean audio and visual data are

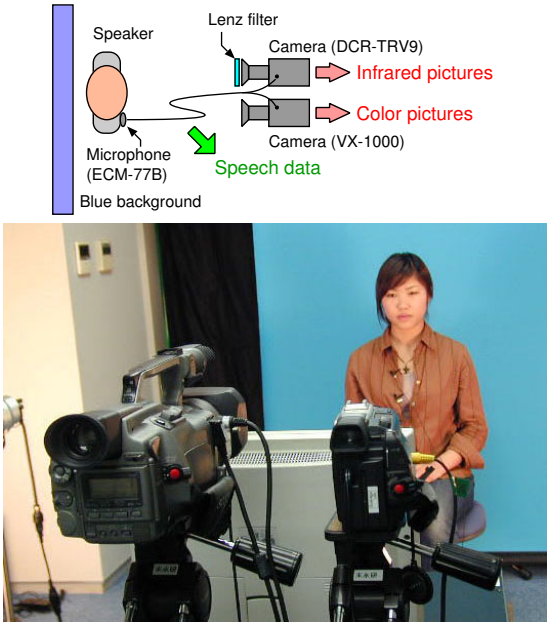
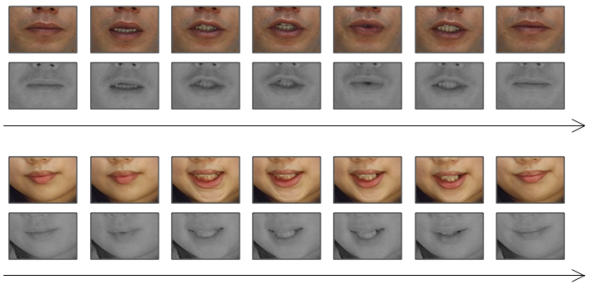


Figure 1: A recording condition for CENSREC-1-AV.



upper: color pictures, lower: infrared pictures

Figure 2: Color and infrared images in CENSREC-1-AV.

employed in the training set.

The sampling frequency of audio signals was converted into 16kHz, subsequently speech data were extracted by Voice Activity Detection (VAD). There were approximately 800ms non-speech signals before and after each detected speech section. Finally the speech data were stored in Microsoft WAV files. Regarding color and infrared movies, each movie was respectively divided into non-interlaced still pictures keeping the aspect ratio. An extraction window was set in order to obtain mouth images. The size of the window was determined as 81×55 for all utterances, and the position was set for each utterance semi-manually (using automatic mouth detection results [12, 13]). Infrared mouth pictures were converted into 8bit gray-scale images, whereas color pictures remained 24bit RGB. Every pictures were stored in Microsoft BMP files. Sample color and infrared images were shown in Figure 2.

2.3. Testing data

A testing data set consists of 1,963 utterances spoken by 26 female and 25 male subjects all who did not participate in the training set. Not only clean audio and visual data but also noisy data were included in the testing set.

In-car noises recorded on city roads and expressways were respectively added to clean speech data at several SNR levels

Table 1: Training and testing sets in CENSREC-1-AV.

	Training set	Test set
# spkr.	22 males and 20 females	25 males and 26 females
# utter.	3,234 utterances	1,963 utterances
Speech data	clean	clean, in-car noise (city roads, expressways)
	monaural, 16kHz, 16bit, WAV files	
Image data (lip-around pictures)	clean	clean, simulated data (Gamma-controlled)
	81×55, 29.97fps, BMP files 24bit RGB (color), 8bit grayscale (infrared)	

(20dB, 15dB, 10dB, 5dB, 0dB and -5dB). Visual distortion was also conducted by simulating a driving-car condition. A gamma value $\gamma(t)$ at a time t was obtained using an intensity value $I(t)$ in a driving car as:

$$\gamma(t) = \frac{\log I(t) - \log 255}{\log \bar{I} - \log 255} \quad (1)$$

where $\bar{I}$ is the average of $I(t)$. Then an intensity value $I(x, y, t)$ at a point (x, y) in an image at a time t was modified as:

$$I'(x, y, t) = 255 \left\{ \frac{I(x, y, t)}{255} \right\}^{\gamma(t+\tau)} \quad (2)$$

where τ is a beginning index determined for each utterance. Figure 3 indicates a sample of distorted pictures. In this example, gamma values recorded across an overpass were used. A mouth is not found in several pictures, or sometimes it becomes unclear. Using the distorted pictures, robustness of a user's visual feature extraction method can be verified. Finally Table 1 summarizes the specification of the CENSREC-1-AV database.

3. Baseline system

3.1. Summary

CENSREC-1-AV contains a basic bimodal speech recognition method as a baseline system. The baseline method is useful as a starter kit of bimodal ASR. By comparing a user's bimodal ASR scheme with the baseline, the user can evaluate the scheme in the common framework. In addition, a user can concentrate efforts on a particular module, e.g. visual feature extraction since the user don't need to develop the other modules.

In the baseline system, audio and visual feature extraction, audio-visual feature integration, audio and visual model training and recognition modules are provided. All programs and scripts are designed to run on Windows and Linux platforms. Note that HTK [11] is required to perform the baseline.

3.2. Audio feature

In each frame, 12-dimensional Mel Frequency Cepstral Coefficients (MFCCs) and a static log power, as well as their first and second derivatives are extracted. As a result, a 39-dimensional audio feature is obtained at every 10ms.

3.3. Visual feature

In CENSREC-1-AV, two kinds of visual features are prepared: a 30-dimensional eigenface parameter [3] or a 6-dimensional optical-flow feature [2]. Since the frame rate is 29.97Hz which

upper: color pictures, lower: infrared pictures

Figure 3: Distorted pictures by the gamma correction technique.

Figure 4: Eigenvectors for color pictures.

is different from that of audio parameters, feature interpolation is subsequently conducted using a 3-degree spline function in order to make the feature rate 100Hz.

3.3.1. Eigenface feature

In order to calculate eigenface visual features, eigenvectors are required beforehand. After 4,620 pictures are extracted from the training data set, downsizing and gray-scale conversion are applied to the images in order to obtain thumbnails of which size are 40×27. Each thumbnail picture is then transformed to a 1,040-dimensional single vector. Principle Component Analysis (PCA) is conducted using all the training vectors, resulting 10 eigenvectors corresponding to the most largest 10 eigenvalues. Figure 4 shows reformed eigenvectors for color pictures. Now using the eigenvectors, component scores of each picture in the training and testing data sets can be computed, then a 10-dimensional static score vector is generated. A 30-dimensional visual feature is finally obtained after calculating first and second derivatives of the static vector.

3.3.2. Optical-flow feature

An optical-flow features are also provided in the baseline to recognize visual speech. It is also effective especially to detect speech and non-speech sections; identifying speech sections increases lipreading performance in noisy conditions. At first, optical-flow vectors are calculated using neighbor two pictures. In this baseline system, we employ the Horn-Schunck method [14] to obtain the flow vectors. Secondly, horizontal and vertical variances of flow vectors at all points are computed. Their delta and delta-delta parameters are thirdly obtained, resulting a 6-dimensional visual feature.

3.4. Audio-visual feature

An audio feature and an interpolated visual feature is simply concatenated to generate an audio-visual feature frame by frame. For example, 69-dimensional audio-visual vectors are obtained when using eigenface parameters.

3.5. Model training and recognition

Multi-stream Hidden Markov Models (HMMs) are built using training data, then evaluated using a testing set. Similar to CENSREC-1, 11 digit HMMs, a silence HMM, and an additional short-pause HMM are employed. All HMMs are left-to-right, and the numbers of states are 16, 3, and 1 respectively. The central state in the silence HMM and the state in the short-pause HMM are shared (tied state). The instructions to build each multi-stream HMM in the baseline method is described as follows:.

(1) Audio HMMs are built using audio features by the embedded training algorithm. This procedure is the same as a conventional ASR.

(2) Using the audio HMMs and the training audio features, a forced alignment technique is conducted to obtain a time-aligned transcription of the training data.

(3) Initial visual HMMs are prepared. A transition matrix in each visual HMM is copied from that in the corresponding audio HMM.

(4) Visual HMMs are built by applying the Baum-Welch re-estimation method, the time-aligned label and visual features. Mean and variance values in every Gaussian components as well as mixture weights are updated while a transition matrix remains.

(5) Using audio model parameters and a transition matrix in the audio HMM as well as visual parameters in the visual HMM, a multi-stream HMM is generated for each digit and silence model.

A multi-stream HMM has two streams (an audio stream and a visual stream) and stream weight parameters called an audio stream weight λ_A and an visual stream weight λ_V. A log likelihood of an audio-visual feature $b_{AV}(\mathbf{o}_{AV})$ is represented as:

$$b_{AV}(\mathbf{o}_{AV}) = \lambda_A b_A(\mathbf{o}_A) + \lambda_V b_V(\mathbf{o}_V) \tag{3}$$

where $b_A(\mathbf{o}_A)$ and $b_V(\mathbf{o}_V)$ are log likelihoods in audio and visual streams respectively. The stream weights are restricted by the following equation:

$$\lambda_A + \lambda_V = 1 \tag{4}$$

When recognizing test data, the same stream weight values are used for all multi-stream HMMs. After 11 candidate weight values (0.0, 0.1, $\cdots$, 0.9, 1.0) are tested, the stream weights that performs the highest accuracy are employed for each audio noise and visual distortion conditions.

Recognition results of audio-only (using the model made by (1) in the training procedure), visual-only (using the training algorithm denoted in (1) and visual features), and bimodal ASRs are enumerated in an evaluation spreadsheet. Using the sheet, a user can compare a user's proposed method with the common baseline and discuss the results.

4. Experiments

Using the database and the baseline system, a recognition experiment is conducted to evaluate the performance of the baseline. In the experiment, the numbers of mixture components in

Table 2: Experimental results of the baseline system using noisy audio and clean visual data.

(A) city-road noise

SNR	Audio only	Bimodal ASR $\lambda_A = 1$	Bimodal ASR best (λ_A)	Bimodal ASR $\lambda_A = 0$	Visual only
-5dB	80.92	80.92	87.73 (0.7)		
0dB	93.34	93.34	95.67 (0.8)		
5dB	98.22	98.22	98.76 (0.8)		
10dB	99.18	99.18	99.32 (0.8)	35.08	27.58
15dB	99.41	99.41	99.50 (0.8)		
20dB	99.60	99.60	99.63 (0.8)		

(B) expressway noise

SNR	Audio only	Bimodal ASR $\lambda_A = 1$	Bimodal ASR best (λ_A)	Bimodal ASR $\lambda_A = 0$	Visual only
-5dB	58.05	58.05	72.28 (0.7)		
0dB	85.42	85.42	90.49 (0.8)		
5dB	97.05	97.05	98.23 (0.8)		
10dB	99.07	99.07	99.26 (0.8)	35.08	27.58
15dB	99.43	99.43	99.52 (0.9)		
20dB	99.66	99.66	99.66 (1.0)		

multi-stream HMMs are the same as CENSREC-1: 20 for digit HMMs and 36 for the other HMMs.

Table 2 shows the recognition accuracy of the baseline for noise-added speech and color pictures. Audio-only and visual only results are obtained by building unimodal HMMs using the embedded training and testing the model respectively. In the bimodal best recognition result, not only the accuracy but also the audio stream weight value are denoted.

According to Table 2, the audio-only and bimodal ($\lambda_A = 1$) results were almost the same since all parameters in the audio stream derive from the audio-only model. In all conditions except SNR-20dB expressway noise, the bimodal ASR achieved significant improvements from the audio-only and visual-only results. Particularly at -5dB conditions, roughly 35% relative error reductions were observed compared to audio-only results. It is also found that an audio stream weight value became small in low SNR conditions. Visual information played an important role in such the noise environments, compensating distorted audio information and improving recognition results. It is notable that the bimodal speech recognition result ($\lambda_A = 0$) was superior to the visual-only result. Both methods used only visual information in recognition, whereas in training, bimodal ASR used audio data but visual-only ASR did not. Therefore, it is concluded that employing audio information in training is effective for lipreading in order to improve the performance.

5. Conclusion

This paper introduces our new corpus for bimodal speech recognition CENSREC-1-AV. The database consists of training and testing data sets, each including audio and visual speech data. As an evaluation framework, the common baseline method using acoustic and visual (eigenface or optical-flow) features as well as multi-stream HMMs is provided. A recognition experiment was conducted to evaluate bimodal speech recognition, then the baseline system achieved significant improvements compared with audio-only and visual-only schemes. Our future work includes the investigation of effective bimodal ASR (features, training and model optimization) by using this corpus, and the construction of a real-environment audio-visual corpus CENSREC-2-AV.

6. Acknowledgement

The authors would like to thank Suenaga laboratory in Nagoya University for their cooperation, and Speech Resource Consortium in National Institute of Informatics for their support.

7. References

[1] G.Potamianos et al., "Discriminative training of HMM stream exponents for audio-visual speech recognition," Proc. ICASSP'98, vol.6, pp.3733-3736 (1998).

[2] K.Iwano et al., "Bimodal speech recognition using lip movement measured by optical-flow analysis," Proc. HSC2001, pp.187-190 (2001).

[3] C.Miyajima et al., "Audio-visual speech recognition using MCE-based HMMs and model-dependent stream weights," Proc. ICSLP 2000, vol.2, pp.1023-1026 (2000).

[4] Y.Lan et al., "Comparing visual features for lipreading," Proc. AVSP2009, pp.102-106 (2009).

[5] S.Nakamura et al., "State synchronous modeling of audio-visual information for bi-modal speech recognition," Proc. ASRU2001 (2001).

[6] S.Tamura et al., "A stream-weight optimization method for audio-visual speech recognition using multi-stream HMMs," Proc. ICASSP2004, vol.1, pp.857-860 (2004).

[7] P.Shen et al., "Evaluation of real-time audio-visual speech recognition," Proc. AVSP2010 (2010).

[8] S.Pigeon et al., "The M2VTS multimodal face database (Release 1.00)," Audio- and Video-based Biometric Person Authentication, Springer, vol.1206, pp.403-409 (1997).

[9] G.Potamianos et al., "Speaker independent audio-visual database for bimodal ASR," Proc. AVSP'97, pp.65-68 (1997).

[10] S.Nakamura et al., "AURORA-2J: An evaluation framework for Japanese noisy speech recognition," IEICE Trans. on Information and Systems, Vol.E88-D, No.3, pp.535–544 (2005).

[11] `http://htk.eng.cam.ac.uk/` .

[12] `http://www.omron.com/r_d/coretech/vision/okao.html` .

[13] S.Tamura et al, "Improvement of audio-visual speech recognition in cars," Proc. ICA2004, vol.4, pp.2595-2598 (2004).

[14] B.K.P.Horn et al., "Determining optical flow," Artificial Intelligence, vol.17, pp.185-203 (1981).