CENSREC-AV: Evaluation frameworks for audio-visual speech recognition

Satoshi TAMURA[1], *Chiyomi MIYAJIMA*[2], *Norihide KITAOKA*[2],
Satoru HAYAMIZU[1], *and Kazuya TAKEDA*[2]

[1] Department of Information Science, Gifu University, Japan
[2] Graduate School of Information Science, Nagoya University, Japan
tamura@info.gifu-u.ac.jp, miyajima@is.nagoya-u.ac.jp, kitaoka@nagoya-u.jp,
hayamizu@gifu-u.ac.jp, kazuya.takeda@nagoya-u.ac.jp

Abstract

This paper introduces incoming evaluation frameworks for bimodal speech recognition in noisy conditions and real environments. In order to develop a robust speech recognition in noisy environments, bimodal speech recognition which uses acoustic and visual information has been paid attention to particularly for this decade. As a lot of methods and techniques for bimodal speech recognition have been proposed, a common evaluation framework, including audio-visual speech data and baseline system, is needed to estimate and compare these techniques and bimodal speech recognition schemes. Audio-visual evaluation frameworks, CENSREC-1-AV and CENSREC-2-AV, have been being built by the CENSREC project in Japan; CENSREC-1-AV includes artificially noise-added waveforms and image sequences, whereas CENSREC-2-AV consists of audio-visual data recorded in in-car environments. A baseline method and its recognition results will be also provided with these corpora.

Index Terms: evaluation framework, audio-visual speech corpus, bimodal speech recognition, noisy environments.

1. Introduction

For mobile devices such as laptops, cell phones and personal digital assistants (PDAs), speech recognition has the potential to become a smart input interface. For example, in in-car conditions, a hands-free interface is strongly required to operate an automotive navigation system or cell phones. However, there has been a serious problem in speech recognition; the performance of speech recognition is drastically degraded in noisy conditions or in real environments.

Since visual information, such as lip shapes and movements, is not affected by acoustic noises, bimodal (or audio-visual, multimodal) speech recognition has attracted attention as a method to ensure the robustness of speech recognition. To investigate bimodal speech recognition, audio-visual speech databases have already been made by many researchers, where a number of speakers uttered digits, words, and sentences. On the other hand, an evaluation framework for bimodal speech recognition, which includes not only audio-visual speech data but also test scripts (baseline results), is not common. Such a framework is essential to develop an effective visual feature, an image processing method, an audio-visual fusion technique, and finally a bimodal speech recognition system.

In this paper, we describe our recent efforts to build new evaluation frameworks for bimodal speech recognition, CENSREC-1-AV and CENSREC-2-AV; these databases are available for robust audio-visual speech recognition in acoustically and visually noisy environments.

Table 1: Typical audio-visual speech databases.

DB name	Audio	Video	spk.	Task
Tulips1 (1995) [E]	clean	only lip region	12	four digits
DAVID (1996) [E]	clean	complex	123	10-digit numbers, alphabets, and sentences
M2VTS (1998) [F]	clean	gray back	37	a 10-digit number
XM2VTS (1999) [E]	clean	blue back	295	10-digit numbers, and a sentence
M2TINIT (2001) [J]	clean	blue back	1	503 phonetically- balanced sentences
AVICAR (2004) [E]	noisy in car	passen- ger seat	100	digits, alphabets, and sentences

spk. = # speakers,
[E] = English, [F] = French, [J] = Japanese.

This paper is organized as follows: current audio-visual speech databases are investigated in Section 2. Section 3 describes the CENSREC project in brief. In Section 4 our frameworks CENSREC-1-AV and CENSREC-2-AV are introduced. Finally Section 5 concludes this paper.

2. Audio-visual speech databases

Bimodal speech recognition has been investigated particularly since the 1980s. Therefore, many audio-visual databases are currently available. Table 1 shows typical audio-visual databases. Furthermore, some researchers built larger or more useful audio-visual databases for their own works, e.g. speaker-independent, LVCSR, and profile-image bimodal speech recognition methods [1, 2].

3. CENSREC project

A working group of the Corpora and Environments for Noisy Speech RECognition (CENSREC) has been established in the Information Processing Society of Japan (IPSJ). For these five years, the CENSREC group has collected numerous amount of speech data in different environments and purposes, resulting five Japanese corpora shown in Figure 1. Table 2 represents a summary of existing corpora as well as incoming audio-visual corpora described in the following section. Researchers can utilize these databases without any fee for research use. More in-

Table 2: A summary of existing and incoming CENSREC corpora.

DB name	release	task	recognition condition
CENSREC-1	7/2003	connected digit	simulated noisy recognition
CENSREC-2	11/2005	connected digit	real-environment recognition
CENSREC-3	2/2005	isolated word	real-environment recognition
CENSREC-1-C	9/2006	connected digit	VAD in noisy environments
CENSREC-4	3/2008	connected digit	reverberant speech recognition
CENSREC-1-AV	coming	connected digit	simulated noisy audio-visual recognition
CENSREC-2-AV	coming	connected digit	real-environment audio-visual recognition

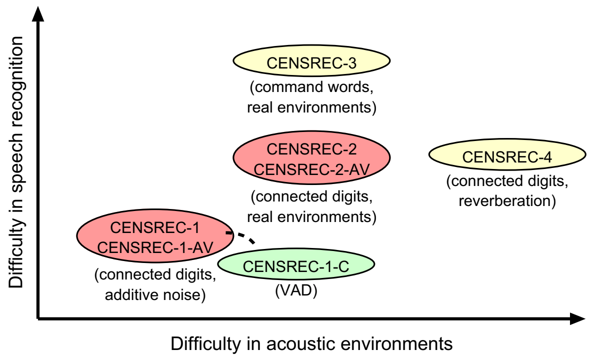

Difficulty in speech recognition
CENSREC-3
(command words, real environments)
CENSREC-2
CENSREC-2-AV
(connected digits, real environments)
CENSREC-4
(connected digits, reverberation)
CENSREC-1
CENSREC-1-AV
(connected digits, additive noise)
CENSREC-1-C
(VAD)
Difficulty in acoustic environments

Figure 1: Difficulties of each corpus in the CENSREC series.

formation about the CENSREC project are described in [3].

3.1. CENSREC-1

Focusing on the effects of additive noises, CENSREC-1 [4], previously called as AURORA-2J, was built as a Japanese version of AURORA-2, a noisy continuous digit recognition database developed in Europe [5].

In this corpus, 55 females and 50 males uttered Japanese connected digit sequences. The length of each sequence is from one to seven. The total number of utterances is the same as AURORA-2. There are two training conditions i.e. clean and multi-condition, whereas the test set consists of three subsets, respectively using a band filter and some kinds of additive noises in five SNR conditions. Training and baseline test scripts based on HTK [6] are also provided, so that a researcher could show one's own results with tables of absolute accuracy and relative improvements from the baseline results.

3.2. CENSREC-2

CENSREC-2 is another database for the evaluation of noisy digit recognition, focusing on real noisy environments [7]. The task is the same as CENSREC-1, but all utterances were recorded in a driving car. There are many recording conditions, e.g. vehicle speed and in-car condition. This database includes 17,651 utterances spoken by 104 speakers (73 for training and 31 for testing). Four evaluation conditions are considered, and a spread sheet is also provided.

3.3. Other existing corpora

CENSREC-3 is an isolated word database recorded in real environments [8]. In this corpus, 50 words were adopted as command words for a navigation system. For research of Voice Activity Detection (VAD), the evaluation framework under noisy environments, CENSREC-1-C, was built in 2006 [9]. This framework consists of noisy continuous digit utterances and evaluation tools for VAD results. To evaluate distant-talking speech recognition in various reverberation environments, CENSREC-4 was constructed and has been distributed [10]. The speech data in this database are connected digit utterances as in CENSREC-1, recorded in eight types of reverberation environments.

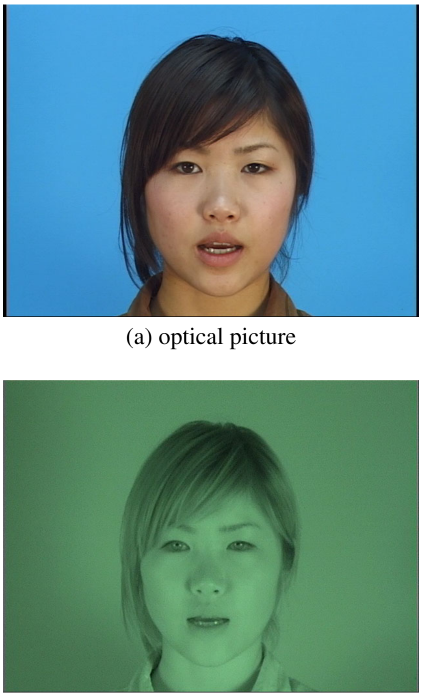
4. CENSREC-AV corpora

As mentioned, bimodal speech recognition is useful to achieve better accuracy in noisy environments. We have been collected and building audio-visual evaluation frameworks: CENSREC-1-AV and CENSREC-2-AV. Both frameworks includes audio-visual databases, that are corresponding to CENSREC-1 and CENSREC-2 respectively, that is, CENSREC-1-AV for additive noises and CENSREC-2-AV for real environments. Regarding video images, not only optical (color) pictures that are commonly used in bimodal speech recognition, but also infrared images are adopted. And we are planning to distribute visually 'noisy' data in both corpora. Through these frameworks, we aim for the following:

1. **Clarification of effective visual features**
   In contrast to acoustic features, no crucial visual feature is found that provides enough and robust visual information of utterances in real environments. We believe that a large-scale audio-visual corpora can statistically establish the effective visual features.
2. **Development of a robust bimodal speech recognition**
   When audio and visual information are incorporated, two kinds of methods are proposed: early integration (feature fusion) and late integration (decision fusion). Using a large-scale databases, these schemes can be estimated or improved in noisy environments.
3. **Modeling of asynchronous information**
   It is widely known that audio-visual asynchronicity exists: for example, when a vowel or a consonant is pronounced, acoustic signals are observed after a speaker's mouth starts opening or moving. In bimodal speech recognition, therefore, the asynchronicity must be properly dealt with. Audio-visual speech database is thus required for training the asynchronicity since neither an

Table 3: Audio specifications of both corpora.

sampling rate	16kHz (downsampled from 48kHz)
quantization bit	16bit
acoustic feature	12 MFCC, 12 Δ MFCC, 12 $\Delta\Delta$ MFCC, power, Δ power, and $\Delta\Delta$ power (39 dimension)


(a) optical picture

(b) infrared picture

Figure 2: Captured pictures for CENSREC-1-AV

audio-only nor a visual-only database can be used for this purpose.

4. **Comparison of recognition results with the common baseline**

   There are many researches regarding bimodal speech recognition, however, these methods cannot be compared and estimated since database and experimental condition are different. To evaluate bimodal speech recognition methods, a common framework or a baseline system is needed. CENSREC-1-AV and CENSREC-2-AV provides the common benchmark of bimodal speech recognition, which is so useful for all related works.

Note that both databases will be released in late 2008 or 2009, and the specifications of these final editions may be updated.

4.1. CENSREC-1-AV

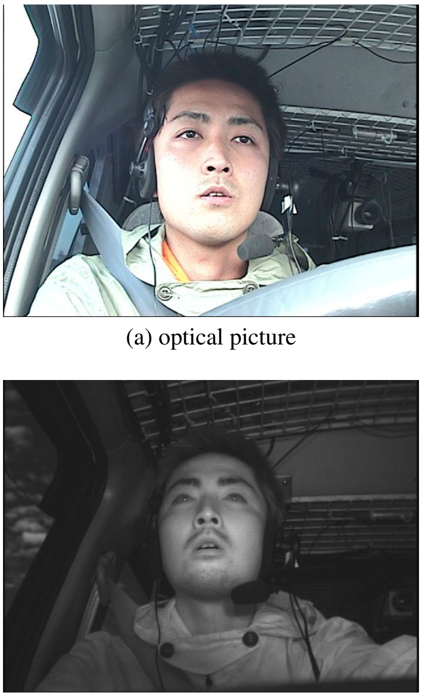
CENSREC-1-AV is a Japanese audio-visual speech database, similar to CENSREC-1 corpus.

Speech data were recorded in acoustically and visually clean conditions. In a recording room where a subject uttered connected digits on a blue background, speech signals were recorded on a lapel microphone, and face movies were captured using optical and infrared cameras. Figure 2 shows a sample picture of each movie. After temporal synchronization of both videos, mouth picture sequences were then extracted using the OKAO-VISION [11] and an HMM-based mouth detection method [12]. As a result, a speech signal and corresponding optical and infrared mouth image sequences can be used for each digit sequence. Audio and visual specifications are shown in Tables 3 and 4, respectively.

CENSREC-1-AV consists of both training and test data sets: 43 subjects (3,311 digit sequences in total) for training, and 51 subjects (1,963 sequences) for testing. The test set consists of not only acoustically and visually clean speech but also noise-adding data; for example, auditory in-car noises in different SNR levels as well as visual intensity noises. In order to estimate the baseline performance, principle-component-analysis (PCA) scores [13] and optical-flow parameters [14] will be provided as conventional visual features. The baseline result using the acoustic and visual features will be made for early-integration and late-integration methods, respectively. Therefore, the user can show the comparison result of proposed and baseline visual parameters, and also, the user can discuss the fusion methods including early and late fusion schemes.

Table 4: Visual specifications of both corpora.

frame rate	30Hz (interlaced movie)
original image size	320x240
object of shooting	mouth/lip
visual feature (for baseline)	PCA score (80% cumulative energy), and/or optical-flow parameter


(a) optical picture

(b) infrared picture

Figure 3: Captured pictures for CENSREC-2-AV.

Table 5: Specifications of CENSREC-1-AV and CENSREC-2-AV.

	CENSREC-1-AV	CENSREC-2-AV
major purpose	model training	test set
place	recording studio	driver's seat in a car
audio condition	clean	noisy
microphone	lapel microphone	microphone on dashboard
video condition	blue background	complicated
camera	optical, infrared	optical, infrared
task	connected digits	connected digits
language	Japanese	Japanese
# subjects for training	20 females and 23 males	—
# subjects for testing	26 females and 25 males	27 females and 30 males
# sequences in total	5,274	6,869

4.2. CENSREC-2-AV

Following CENSREC-1-AV, another database CENSREC-2-AV will be also distributed. This database aims at evaluating bimodal speech recognition in real-world environments, corresponding to CENSREC-2.

A Toyota Hiace Regius minivan was used to record in-car audio-visual speech data. Optical and infrared cameras were respectively located on the dashboard and on the steering column in front of a driver's seat, as well as a hands-free microphone. Subjects sat in the driver's seat and uttered connected digit sequences as in CENSREC-2, while operating the vehicle on inner-city roads. In this database, audio-visual speech data made by 56 subjects will be included. Figure 3 shows pictures captured by both cameras. Similar to CENSREC-1-AV, pictures around subject mouths were edited out of the optical and infrared movies. Note that a speaker's mouth was sometimes hidden by her/his arm or the steering wheel; furthermore, lighting conditions frequently and greatly changed with shadow of buildings since the car drove on a city road. In the audio channel, speech signals were disturbed by driving noises, resulting in a low SNR value. CENSREC-2-AV thus provides much challenging task to bimodal speech recognition.

In some cases, CENSREC-2-AV may be employed for testing a bimodal speech recognition system whose acoustic and visual recognition models are built using CENSREC-1-AV. Therefore, the baseline result will be provided using the audio-visual parameters as well as a model made using these features of CENSREC-1-AV data, and the speech data in CENSREC-2-AV as a test set. Table 5 compares the specifications of CENSREC-1-AV with CENSREC-2-AV.

5. Conclusion

This paper describes our evaluation framework, CENSREC-1-AV and CENSREC-2-AV. CENSREC-1-AV includes audio-visual speech data recorded in clean condition as well as acoustically and visually noise-added data, whereas CENSREC-2-AV consists of audio-visual data in real in-car environments. All speech data have already been recorded, then we are now preparing to distribute these frameworks.

Our future work includes: (1) building up and release of both databases (maybe started in 2009), (2) research for effective bimodal speech recognition using the corpora, and (3) construction of a next corpus (e.g. CENSREC-3-AV).

6. Acknowledgments

The authors would like to thank the other members of CENSREC working team for their assistance about our work.

7. References

[1] G.Potamianos et al., "Hierarchical discriminant features for audio-visual LVCSR," Proc. ICASSP2001, vol.1, pp.165-168 (2001-5).

[2] T.Yoshinaga et al., "Audio-visual speech recognition using lip movement extracted from side-face images," Proc. AVSP2003, pp.117-120 (2003-9).

[3] http://sp.shinshu-u.ac.jp/CENSREC/index.html.en .

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

[5] H.Hirsh et al., "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions," Proc. ASR2000, pp.175-180 (2000-9).

[6] http://htk.eng.cam.ac.uk/.

[7] S.Nakamura et al., "CENSREC-2: Corpus and evaluation environments for in-car continuous digit speech recognition," Proc. INTERSPEECH2006, pp.2330-2333 (2006-9).

[8] M.Fujimoto et al., "CENSREC-3: An evaluation framework for Japanese speech recognition in real friving-car environments," IEICE Transactions on Information and Systems, vol.E89-D, no.11, pp.2783-2793 (2006-11).

[9] N.Kitaoka et al., "CENSREC-1-C: An ealuation framework for voice activity detection under noisy environment," Proc. ASRU2007, pp.607-612 (2007-12).

[10] T.Nishiura et al., "Evaluation framework for distant-talking speech recognition under reverberant environments – Newest part of the CENSREC series –," Proc. LREC2008 (2008-5).

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

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

[13] T.Togo et al., "Comparison of visual features for audio-visual speech recognition using the AURORA-2J-AV database," Proc. ASA & ASJ Joint Meeting, p.3044 (2006-11).

[14] K.Iwano et al., "Bimodal speech recognition using lip movement measured by optical-flow analysis," Proc. HSC2001, pp.187-190 (2001-4).