FUSION OF STANDARD AND ALTERNATIVE ACOUSTIC SENSORS FOR ROBUST AUTOMATIC SPEECH RECOGNITION

Panikos Heracleous, Jani Even, Carlos T. Ishi, Takahiro Miyashita, and Norihiro Hagita

ATR, Intelligent Robotics and Communication Laboratories, Japan

{panikos,even,carlos,miyasita,hagita}@atr.jp

ABSTRACT

This paper focuses on the problem of environmental noises in human-human communication and in automatic speech recognition. To deal with this problem, the use of alternative acoustic sensors –which are attached to the talker and receive the uttered speech through skin or bones– is investigated. In the current study, throat microphones and ear bone microphones are integrated with standard microphones using several fusion methods. The results obtained show that the recognition rates in noisy environments are drastically increased when these sensors are integrated with standard microphones. Moreover, the system does not show any recognition degradations in clean environments. In fact, recognition rates also increase slightly in clean environments. Using late fusion to integrate a throat microphone, an ear bone microphone, and a standard microphone, we achieved a 44% relative improvement in recognition rate in a noisy environment and a 24% relative improvement in recognition rate in a clean environment.

Index Terms— Alternative sensors, ear bone microphone, throat microphone, fusion, robust speech recognition

1. INTRODUCTION

In human-human communication and in automatic speech recognition robustness against environmental noises is a critical issue. Although automatic speech recognition in clean environments shows very high recognition rates, the performance of speech recognition systems operating in noisy environments drastically decreases. Currently, several methods deal with this problem. These methods include speech de-noising, model adaptation to noisy conditions, multimodal processing, and use of microphone arrays.

In addition to these methods, several studies have been introduced which use alternative sensors to capture uttered speech [1, 2]. These sensors are attached to the talker and receive the uttered speech directly through skin and bone. As a result, alternative sensors show higher robustness against noise compared with air-conductive microphones. The majority of the related studies use throat microphones [3] or a combination with standard microphones [4, 5, 6] to capture uttered speech.

This work was supported by KAKENHI (21118001).

(a) (b)

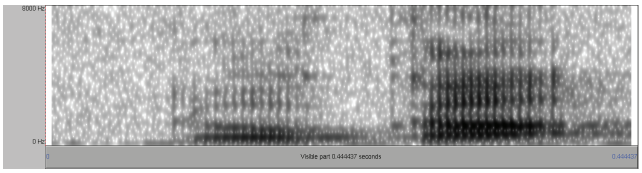
Fig. 1: (a) Throat microphone (b) Ear bone microphone.

The main difference between the current work and the previous works is the use of a new ear bone microphone as an additional alternative sensor and also the methods used to integrate several acoustic sensors. In the current work, however, a feature fusion method and a late fusion method were used to integrate the sensors. In addition, in the case of late fusion, an adaptive weighting method was proposed, which does not require any adjustment of the stream weights. The authors also suggest a method for automatic segmentation of noisy speech data by using an alternative sensor along with the desired microphone during recording.

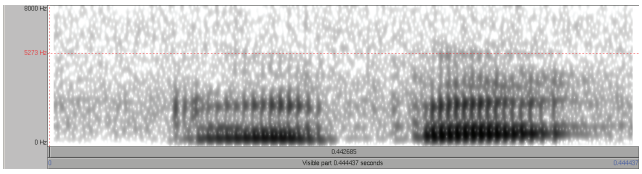
2. ALTERNATIVE ACOUSTIC SENSORS

Figure 1 shows the throat and ear bone microphones used in this study. Both are commercial and inexpensive products. The throat microphone is attached to the speaker's neck; it captures the vibrations through the skin. An ear bone microphone is attached inside the talker's ear and receives the uttered speech through the jaw bone.

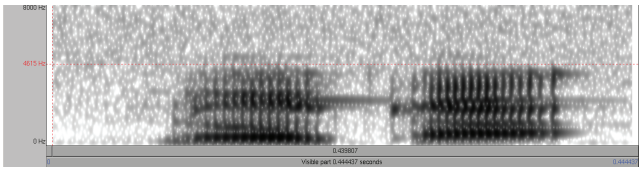
Figure 2 shows the spectrogram of an utterance received by the three sensors. Given that skin and bones act as a low-pass filter, the upper frequencies for both throat and ear bone microphones are not included in the speech signal. In the



(a) Standard microphone



(b) Ear bone microphone



(c) Throat microphone

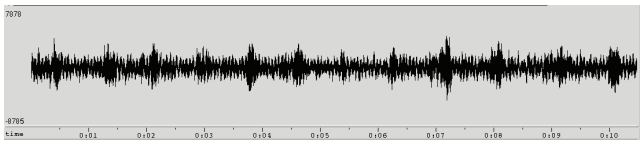
Fig. 2: Spectrogram of a clean utterance received by a standard microphone, an ear bone microphone, and a throat microphone.

case of the throat microphone the upper frequency is about 4600 Hz and in the case of the ear bone microphone the upper frequency was about 5250 Hz. As a result, recognition rates decrease when using an alternative sensor alone in a clean environment, compared with using a standard microphone. On the other hand, in noisy environments alternative sensors show higher recognitions. The current study aims at taking advantage of the different sensors in both clean and noisy environments by integrating them using fusion methods applied in multimodal signal processing [7].

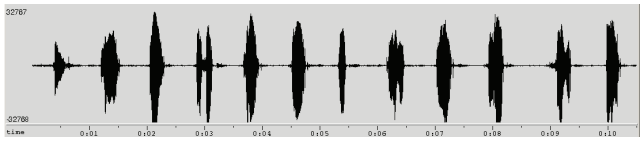
3. METHODOLOGY

3.1. Corpus and statistical modeling

For the automatic speech recognition experiments, two male and two female speakers were employed. The corpus consisted of the 120 words of the Japanese Diagnostic Rhyme Test [8]. Each speaker uttered each word 10 times in a clean environment and 5 times as babble noise at 70 dB(A) was played back through a loudspeaker. For training clean hidden Markov models (HMMs), 5 instances of each word under clean condition were used. For testing, 5 instances of each word under clean conditions and 5 instances of each word under noisy conditions were used. The statistical models were whole-word, 7-state HMMs. Each state was modeled with 2 Gaussian distributions. The feature vectors were of length 36 (i.e., 12 MFCC, 12 ΔMFCC, and 12 $\Delta\Delta$MFCC).



(a) Standard microphone



(b) Throat microphone

Fig. 3: Waveforms of a noisy utterance received by a standard microphone and a throat microphone.

3.2. Data segmentation

For each speaker, the training and test sets were recorded in a single session. Thus before any further processing, it was necessary to segment all of the utterances. Because all of the channels were recorded synchronously, the segmentation task was performed using the channel having the best signal-to-noise ratio (SNR), namely the throat microphone, as shown in Figure 3. In particular, the great immunity of the throat microphone to external noise was very useful for segmenting the noisy part of the data.

Since the SNR for the throat microphone channel was very good, it was possible to segment the data by averaging the power spectrum on a given frequency band ([100, 5000] Hz) and thresholding that average value to detect speech segments. To choose the threshold, we used the structure of our data sets: a quasi periodical succession of short utterances. An iterative algorithm selected the adequate threshold for each of the speakers, so that the proportion of signal samples –the sample above the threshold– was approximately one-third. After this first segmentation, gaps of more than 150 ms and dents of less than 100 ms were suppressed to obtain the final segmentation (head and tail guard intervals of 50 ms were also added). Finally, the quality of the segmentation was assessed by a listening test. In total, 99.53% of the utterances were successfully segmented and a mere 3.95% of extracted segments were noises.

3.3. Integration of multiple sensors

This section introduces the fusion methods used to integrate the several acoustic sensors. In the present study, a feature fusion method and a late fusion methods were used.

3.3.1. Concatenative feature fusion

The feature concatenation is the simplest state synchronous fusion method. It uses the concatenation of two or more sig-

nals as the joint feature vector:

$$O_t^{AB} = [O_t^{(A)^T}, O_t^{(B)^T}]^T \in R^D \quad (1)$$

where, O_t^{AB} is the joint feature vector, $O_t^{(A)}$ is the feature vector of the first sensor, $O_t^{(B)}$ is the feature vector of the second sensor, and D is the dimension of the joint feature vector. In these experiments, the dimension of each stream was 36. Thus, the dimension D of the joint feature vectors was 72 and 108.

3.3.2. Late fusion

In the late fusion method, two single HMM-based classifiers were used for the two sensors. For each test utterance (i.e., isolated word), the two classifiers provided an output list that included all the word hypotheses with their likelihoods. Subsequently, all of the separate mono-modal hypotheses were combined into the bi-modal hypotheses using the weighted likelihoods, as given by:

$$logP_{AB}(h) = \lambda_a logP_A(h|\mathrm{O}_A) + \lambda_b logP_B(h|\mathrm{O}_B) \quad (2)$$

where, $logP_{AB}(h)$ is the score of the combined bi-modal hypothesis h, $logP_A(h|\mathrm{O}_A)$ is the score of the h provided by the first classifier, and $logP_B(h|\mathrm{O}_B)$ is the score of the h provided by the second classifier. λ_a and λ_b are the stream weights with the constraints:

$$0 \leq \{\lambda_a, \lambda_b\} \leq 1, \quad and \quad \lambda_a + \lambda_b = 1 \quad (3)$$

In these experiments, the weights were experimentally adjusted by maximizing the word accuracy on several experiments. In the case of clean test data, the weight of the standard microphone stream was adjusted to 0.7 and the weight of the body-conducted microphone stream was adjusted to 0.3, respectively. In the case of noisy test data, the weight of the standard microphone stream was adjusted to 0.2 and the weight of the body-conducted microphone was adjusted to 0.8, respectively.

The procedure described in this study finally resulted in a combined N-best list in which the top hypothesis was selected as the correct bi-modal output.

3.4. Late fusion with adaptive weights

A disadvantage of the previously described late fusion method is the choice of appropriate weights. Thus, in late fusion, the weights were adjusted experimentally across several experiments, and different weights were used for clean and noisy test data. To avoid this, a novel adaptive weighting method based on the likelihoods of the mono-modal hypotheses in the N-best list is proposed. The weights were self-adjusted by a normalized non-linear transformation of the likelihoods.

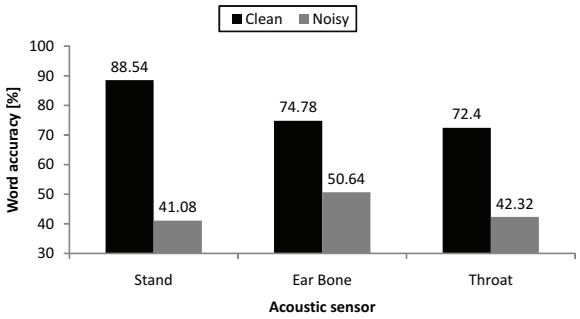


Fig. 4: Word accuracy using clean and real noisy data.

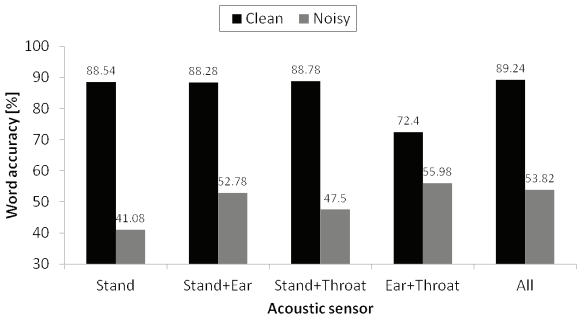


Fig. 5: Word accuracy using clean and real noisy data in the feature fusion method.

In the case of integration of two sensors, the weights will be as follows:

$$\lambda_a = \frac{logP_A(h|\mathrm{O}_A)}{logP_A(h|\mathrm{O}_A) + logP_B(h|\mathrm{O}_B)}$$

and,

$$\lambda_b = \frac{logP_B(h|\mathrm{O}_B)}{logP_A(h|\mathrm{O}_A) + logP_B(h|\mathrm{O}_B)} \quad (4)$$

4. EXPERIMENTS

Figure 4 shows the results obtained when using single sensors. As shown, in the case of clean test speech, the standard microphone achieved the highest word accuracy. In the case of the ear bone and the throat microphones, the word accuracies were similar. Using real noisy test data, the word accuracy drastically decreased in all cases. The highest word accuracy in noisy conditions was obtained when using the ear bone microphone.

Figure 5 shows the results when the sensors used in this study were integrated using concatenative feature fusion. As shown, by integrating the standard microphone with a body-conducted microphone, the word accuracy in noisy conditions increased without any decrease for clean conditions.

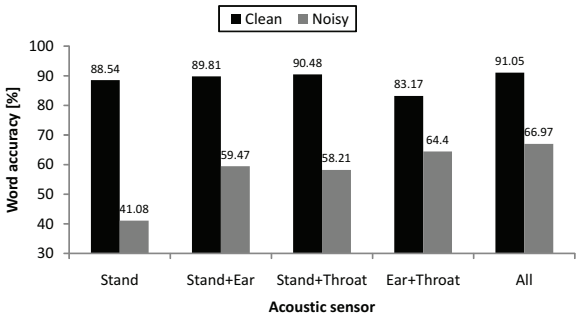


Fig. 6: Word accuracy using clean and real noisy data in the late fusion.

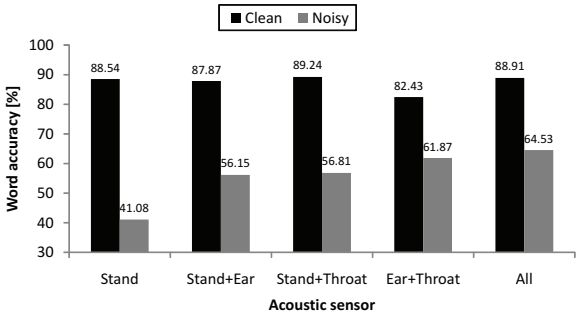


Fig. 7: Word accuracy using clean and real noisy data in the late fusion with adaptive weights.

The highest word accuracy using noisy speech was achieved when integrating ear bone and throat microphones. Using this method, however, the word accuracy for clean data was the lowest one. Integrating all of the three sensors appears to be the most effective solution.

Figure 6 shows the results obtained with the late fusion method. Integrating a body-conducted sensor with the standard microphone, the word accuracy in noisy conditions increased without a decrease for clean conditions. Integrating all of the sensors, the word accuracy increased from 41.08% to 66.97%, with the difference being statistically significant (after conducting the ANOVA test).

Figure 7 shows the results obtained when late fusion with adaptive weighting was used. Using this fusion method, the second best results were achieved. Compared to the late fusion with pre-adjusted weights (i.e., the best case), the differences are not statistically significant. The main advantage is that adjusting the weights for the several streams was not necessary. Also, there was no need to take clean or noisy conditions into account. The weights were self-adjusted based on the likelihoods, and the results show that the proposed method works very well.

5. CONCLUSIONS

In this study, experimental results using alternative acoustic sensors are introduced. Using feature fusion method and late fusion method, a standard microphone was integrated with ear bone microphone and throat microphone resulting in significant improvements in the recognition rates in both clean and noisy environments. A novel adaptive weighting method in late fusion, which does not require any weight adjustments was also introduced.

6. REFERENCES

[1] P. Heracleous, T. Kaino, H. Saruwatari, and K. Shikano, "Unvoiced speech recognition using tissue-conductive acoustic sensor," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, pp. 56, 2007.

[2] O. M. Strand, T. Holter, A. Egeberg, and S. Stensby, "On the feasility of asr in extreme noise using the parat earplug communication terminal," *in Proc. of ASRU*, pp. 315–320, 2003.

[3] S. C. Jou, T. Schultz, and A. Weibel, "Adaptation for soft whisper recognition using a throat microphone," *in Proc. of Interspeech2004-ICSLP*, pp. 5–8, 2004.

[4] Z. Zhang, Z. Liu, M. Sinclair, A. Acero, L. Deng, J. Droppo, X. Huang, and Y. Zheng, "Multi-sensory microphones for robust speech detection,enhancement, and recognition," *in Proc. of ICASSP*, pp. 781–784, 2004.

[5] M. Graciarena, H. Franco, K. Sommez, and H. Bratt, "Combining standard and throat microphones for robust speech recognition," *IEEE Signal Processing Letters*, vol. 10, no. 3, pp. 72–74, 2003.

[6] S. Dupont, C. Ris, and D. Bachelart, "Combined use of close-talk and throat microphones for improved speech recognition under non-stationary background noise," *in Proc. of Robust 2004 (Workshop (ITRW) on robustness issues in conversational interaction*, 2004.

[7] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A.W. Senior, "Recent advances in the automatic recognition of audiovisual speech," *in Proc. of the IEEE*, vol. 91, Issue 9, pp. 1306–1326, 2003.

[8] M. Fujimori, K. Kondo, K. Takano, and K. Nakagawa, "On a revised word-pair list for the japanese intelligibility test," *In Proc. of International Workshop Frontiers in Speech and Hearing Research*, pp. 103–108, 2006.