2009 Fifth International Conference on Information Assurance and Security

Acoustic-based Security System: Towards Robust Understanding of Emergency Shout

Hiroaki NANJO
Faculty of Science and Technology
Ryukoku University
Otsu, Shiga, JAPAN
nanjo@rins.ryukoku.ac.jp

Takanobu NISHIURA, Hiroshi KAWANO
Graduate School of Science and Engineering
Ritsumeikan University
Kusatsu, Shiga, JAPAN
{nishiura@is, cm004042@ed}.ritsumei.ac.jp

***Abstract*—We have been investigating a speech processing system for ensuring safety and security, namely, acoustic-based security system. Focusing on indoor security, we have been studying for an advanced security system which can discriminate emergency shout from the other acoustic sound events based on automatic understanding of speech events. In this paper, we present our investigations, and describe fundamental results.**

***Keywords*-Acoustic-based Security System, Shout, Radiation characteristics, Speech understanding**

I. Introduction

Acoustic-based security system based on understanding of speech events is addressed. Focusing on indoor security, such as school security, we study for an advanced acoustic-based system which can discriminate emergency shout from the other speech events.

Conventional studies of shouted speech have been mainly focusing on speaker detection[1], emotion recognition[2], or event detection[3][4], and have rarely considered making use of linguistic information included in shouts. Actually, in such studies, large sound signals such as screams and gunshots were just detected. At school or home in which there are students and children, large sound signals may occur frequently, that is, they may talk with a loud voice, make delightful shout, or shout for calling their friends. Therefore, for advanced security system, making use of linguistic information is significant in order to discriminate such kind of loud voices from emergency shout. Here, we define two types of emergency speech. One is "scream" which does not contain linguistic information. The other is "shout" which contains it, and thus, is our research target. Automatic speech recognition (ASR), which converts speech signals to texts, is promising for such security system since it helps us to understand speech events and to take appropriate actions. Although ASR works well for polite speech, it does not work well for excited shout which is uttered for help.

Based on the background, we have been studying for robust understanding of emergency shout using ASR and robust shout detection for two years. In this paper, we describe our fundamental results, that is, our shouted speech corpus, characteristics of shout, detection of shout, and ASR of shouted speech.

II. Corpus of shouted speech

First of all, our corpus of shouted speech is described. Here, we constructed the corpus of shouted speech which consists of isolated shouted words. For actual security system, it is significant to investigate what kinds of words are shouted in emergency. According to our discussions, firstly, we selected 50 Japanese words since people are unlikely to shout sentences in emergency. Moreover, it is hard to shout sentences in principle. We recorded shout utterances consisting of 50 Japanese words by 20 male and 20 female speakers (2000 utterances in total). The same speakers normally uttered the same 50 words (2000 natural speeches). All utterances are recorded in anechoic room with three microphones; headset microphone, distant microphones (front / back with 1.0 meters distance). Although we do not use shouted sentences in this work, 20 sentences which are likely to be shouted at railway stations were also recorded.

III. Characteristics of shouted speech

A. Spectral characteristics

Firstly, we have compared shouted speech with natural speech in terms of formant, power and fundamental frequency (F0). Formants of male speakers for natural and shouted speech are listed in Figure 1. In shouted speech, formants are shifted to high frequency domain, and standard deviations are larger. Cepstral distance (mean-squared difference)

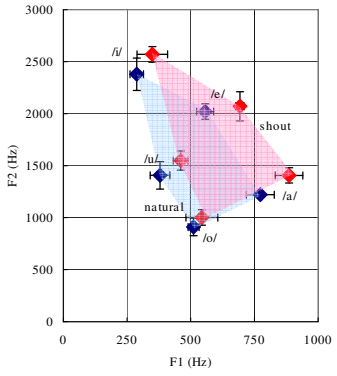

Fig. 1. Formants distributions (9 males)

978-0-7695-3744-3/09 $25.00 © 2009 IEEE
DOI 10.1109/IAS.2009.121

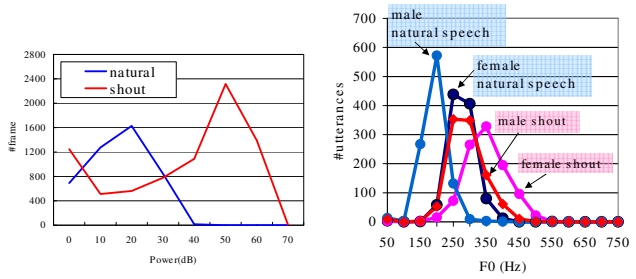

Fig. 2. Power histogram (Left) and F0 histogram (Right)

TABLE I
CEPSTRAL DISTANCE BETWEEN PAIRS OF VOWELS (WITHIN-SPEAKER CEPSTRAL DIST ANCE)

	/i/	/u/	/e/	/o/
/a/	1.19 / 1.17	1.18 / **0.94**	0.97 / **0.79**	0.89 / 0.86
/i/		0.98 / **0.84**	1.23 / **0.90**	1.04 / 1.13
/u/			1.31 / **0.89**	1.04 / 1.07
/e/				1.11 / **0.91**

natural / shouted

between pairs of vowels is listed in Tables I. Here, we used 37 dimensional cepstrum to calculate distances. There is a large cepstral distance between natural and shouted speech. In shouted speech, cepstral distances between /u/, /e/ and other vowels tend to be small. The results show that shouted speech is quite different from natural speech and distinction of vowels would be more difficult.

Power histogram is shown in left side of Figure 2. The result also shows that shouted speech differs from natural speech. Figure 2 also shows F0 histogram. Large overlap between male shouted speech and female natural speech is confirmed. The result shows that gender detection is significant for shout detection[5]. For shout detection, these features are promising.

B. Radiation characteristics

Then, we investigated the radiation characteristics of shouted speech. Here, we recorded five natural and five shouted Japanese vowels in order to measure the radiation characteristic in soundproof room. The recording was synchronously conducted in 15 degrees apartments with 13 microphones. The microphones were located at 1.0 meters distance between microphone and speaker, and 1.5 meters height. We analyzed the radiation characteristics based on the powers. Figure 3 shows the results without energy normalization, and Figures 4 and 5 show the results with normalization. In Figures 3 to 5, 0 degree represents the front direction of the speaker, and 180 degree also represents the back direction. It is confirmed that shouted vowels show bigger energy difference between two microphones located at 1.5 meters distance back and forth than naturally uttered vowels. The result encourages us to detect natural/shouted speech based on above two microphones.

IV. SHOUTED SPEECH DETECTION

For security system, shout detection is significant. Many conventional studies have tried to discriminate voice from the other sound signals, namely, voice activity detection

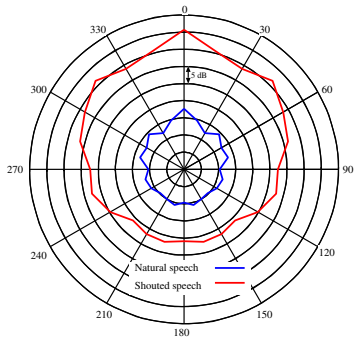

Fig. 3. Radiation characteristics (without energy normalization)

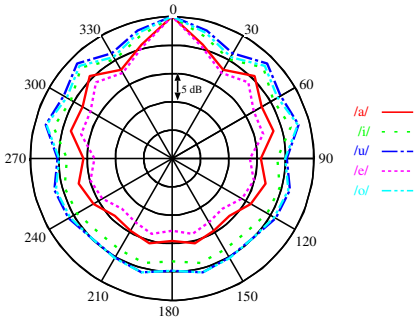

Fig. 4. Radiation characteristic of natural speech (with energy normalization)

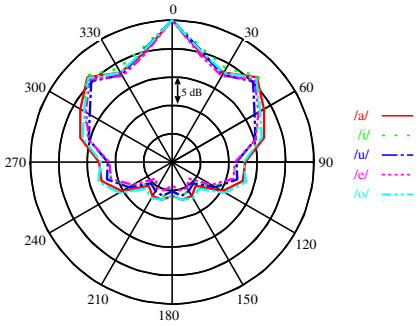

Fig. 5. Radiation characteristic of shouted speech (with energy normalization)

(VAD) [6] [7]. In this paper, we do not perform VAD. We assume that VAD could be perfectly performed, and try to discriminate shout from natural speech.

A. Discrimination based on detected speech power

1) Single microphone case: As shown in Figure 2, speech power seems to be one of the most significant features for discrimination of shout and natural speech. However, speech power (absolute energy) detected by microphone is quite sensitive for distance between microphone and speech source. Figure 6 illustrates speech power of shout detected by distant microphone (1 meters front) and one of natural speech detected by headset microphone. It is confirmed that there is a large overlap of both distributions. In actual security system, we can not control distances between microphone and speaker who is shouting for help. Therefore, discrimination based on absolute energy detected by single microphone is unreliable.

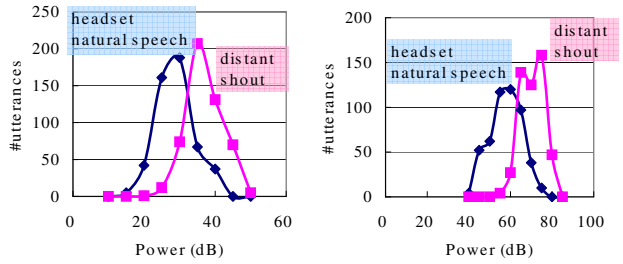

Fig. 6. Distributions of speech power detected by headset and distant microphones (left: averaged power in each utterance, right: max. power of each utterance)

TABLE II
SHOUT DETECTION WITH RADIATION CHARACTERISTICS

Natural Speech	86%
Shouted Speech	92%

TABLE III
SHOUT DETECTION WITH SPECTRAL ENVELOPE FEATURE

microphone	natural → natural (correct)	natural → shout (false alarm)
headset	387 (77.4%)	113 (22.6%)
distant (front, 1m)	333 (66.6%)	167 (33.4%)
roof	323 (64.6%)	177 (35.4%)

We have controlled threshold in the classification so that false rejection (shout → normal) would be 0%.

2) Multiple microphone case – Detection based on radiation characteristics: Next, we have tried to detect shouted speech with multiple microphones. Here, two males were employed, and 50 natural and shouted words for each subject were recorded using two microphones located at 1.5 meters distance back and forth. Equations (1) and (2) show the detection algorithm. In equation (1), $P_f(t)$ represents the captured power for front direction, and $P_b(t)$ also represents one for back direction. In addition, t and N represent time index and frame number of captured signal, respectively. The natural/shouted speeches are detected based on equation (2) with equation (1).

$$e = \sum_{t=1}^{N} (P_f(t)^2 - P_b(t)^2) \quad (1)$$

$$Ident = \begin{cases} Normal\ \ Speech & (e < Th) \\ Shouted\ \ Speech & (e \geq Th) \end{cases} \quad (2)$$

Table II lists the result of detection accuracy. We confirmed that natural/shouted speeches are accurately detected based on their radiation characteristics without conventional speech features such as MFCC (Mel-Frequency Cepstrum Coefficient).

B. Discrimination based on spectral envelope

We have also investigated the discrimination method without absolute speech power. According to our analyses of shouted speech described in section III-A, we adopt an acoustic feature based on MFCC, which models spectral envelope and is commonly used in ASR. Specifications of speech analysis and acoustic feature are described in section V-A. Using the feature, six acoustic models based on continuous density Gaussian-mixture HMMs were trained with shout and natural speech of headset microphone and two distant microphones (1 meters front and roof). For training, we took Maximum Likelihood Linear Regression (MLLR) adaptation. Here, in order to avoid adapting acoustic model to the test speaker, 10-fold cross validation is performed. Specifically, for the classification of utterances of a specific speaker, acoustic model is adapted with the other nine speakers' utterances. The amount of adaptation data is about 10 minutes. For each utterance, acoustic scores (likelihood) are calculated with six HMMs, and then, classification is performed according to the scores. For example, input speech x is determined as "shout" where $P(x|\omega^i_{\text{shout}}) > P(x|\omega^j_{\text{natural}})$. Here, $P(x|\omega^i_{\text{shout}})$ is an acoustic model score (likelihood) derived from HMM trained with shout of i (i = headset, front, roof), and $P(x|\omega^j_{\text{natural}})$ is an acoustic model score derived from HMM trained with natural speech of j (j = headset, front, roof).

The security system should detect all emergency shout. Therefore, false rejection of shout, that is, misclassification of shout to natural speech class is not permitted. We have to evaluate discrimination results from this point of view. In this paper, false alarm (normal → shout) rate under the condition of 0% of false rejection (shout → normal) is used for evaluation. Specifically, we introduced a positive threshold (THRESH) in classification of shout and natural speech, and regarded input speech x as "shout" where $P(x|\omega^i_{\text{shout}}) + \text{THRESH} > P(x|\omega^j_{\text{natural}})$ so that false rejection rate would be 0%.

Table III lists the discrimination results. We confirmed that the false alarm rate is 22.6% to 35.4%, and increases according to the distance between speaker and microphone. The fact shows that we can achieve about 30% of false alarm rate when microphones are placed so that the distance between person and one microphone at least is smaller than 1m. Moreover, since almost all input speech to the actually operating security system is expected to be natural speech, the discrimination method can reduce the number of speech events to be checked by human security officer about 60 to 80%.

For more accurate detection, we have been investigating the use of other features such as radiation characteristics described in section IV-A2 (see also [8]), prolonged vowels in shout (e.g. "pleeeeeeeese")[9], spectral movement, and so on.

V. SHOUTED SPEECH RECOGNITION

A. Effects of model adaptation and vocabulary size

To understand speech events, ASR is indispensable. In this paper, isolated word recognition system is set up with three states left-to-right HMM acoustic model and a decoder Julius [10]. As for acoustic model, gender independent triphone model (2000 states, 16 mixtures) trained with JNAS corpus (normally uttered read speech) is used. Speech analysis is performed every 10 msec., and a 25 dimensional parameter is computed (12 MFCC + 12 Δ MFCC + Δ Power). We used 3 lexicons with vocabulary size of 50, 300, and 500. There are no out-of-vocabulary (OOV) words in the lexicons.

TABLE IV
SHOUTED SPEECH RECOGNITION RESULT (WACC. %)

vocab. size	baseline	MLLR	MAP
50	69.8	94.2	94.2
300	61.2	93.6	92.8
500	60.6	95.0	93.6

test speech: shout recorded with headset microphone

For shouted speech recorded with headset microphone, ASR is performed using baseline acoustic model, which is trained with naturally uttered speech. The results are listed in Table IV ("baseline"). Even for the small vocabulary case, ASR accuracy is 70%, which is not sufficient for security system. In contrast, ASR for naturally uttered speech achieves more than 95%. The results show that shouted speech is quite different from natural speech, and we should recognize shout more accurately.

Then, we have investigated the adaptation of acoustic model to "shouted speech", that is, environmental adaptation. In this work, we took MLLR and Maximum a posteriori Probability (MAP) for acoustic model adaptation. Also here, 10-fold cross validation is performed to avoid adapting acoustic model to the test speaker. The results using an acoustic model adapted to shouted speech by MLLR and MAP methods are also listed in Table IV. Because of small data size, MLLR outperformed MAP in this experiment, and we achieved about 95% of accuracy for both adaptation methods even for 500 word lexicon. The lexicon size must be sufficient for security system. The results shows that if we capture shouted speech under an ideal condition (headset microphone), we can achieve practical accuracy of shouted speech recognition for security system.

B. Speech recognition under several environments

Then, shouted speech recognition under several environments has been investigated. Specifically, we tested with speeches captured with distant microphones and speeches recorded at echo rooms. For recognition, an acoustic model is adapted to corresponding environment. Lexicon size was 50. The results are listed in Table V.

For speech recorded in anechoic room, we achieved about 94% of ASR accuracy even for shout. It is confirmed that the distance between speaker and microphone does not affect shouted speech recognition. For speech recorded in real environments (echo room), ASR accuracy got lower according to the reverberation time. Note that, for natural speech of longer reverberation room, we achieved about 90% of accuracy. Shouted speech recognition is more sensitive for reverberation time than ASR of natural speech. In echo room, reverberated sound signals of prolonged vowels which have stronger power add to succeeding speech signals, and that is one of the main reasons of ASR degradation for shout in longer reverberation rooms. Another possible reason is that MLLR adaptation might not work well since shouts recorded in longer reverberation room are quite different from acoustic model which was trained with speech recorded in anechoic room.

TABLE V
ASR RESULTS UNDER SEVERAL ENVIRONMENTS

environment	channel	natural	shout
anechoic room	headset	96.0%	94.2%
anechoic room	distant (front, 1m)	97.0%	93.6%
anechoic room	roof	94.8%	94.2%
echo room (reverb. 0.3s)	headset	95.6%	90.6%
echo room (reverb. 0.47s)	headset	93.2%	72.0%
echo room (reverb. 0.78s)	headset	89.4%	65.4%

Vocabulary size: 50 (no OOV words).

For each set, MLLR is performed. (10-fold cross validation)

For more accurate recognition, we have been investigating 1) ML training of acoustic model using large corpus of shout, 2) design of HMM topology which models prolonged vowels in shout, and 3) robust features for shout recognition in longer reverberation condition.

VI. CONCLUSION

Acoustic-based security system detecting emergency shout was described. The fundamental results of shouted speech, specifically, characteristics of shouted speech, detection of shout, and ASR results were described. It is confirmed that shouted speech is quite different from natural speech. We also confirmed the significance of further investigations under real reverberation environment including speech analysis, shout detection, acoustic modeling unit, and HMM topology.

ACKNOWLEDGMENT

The authors would like to thank to our laboratory members in Ryukoku University and Ritsumeikan University.

REFERENCES

[1] I. Shahin, "Improving speaker identification performance under the shouted talking condition using the second-order hidden markov models," *EURASIP J. Appl. Signal Process.*, vol. 2005, pp. 482–486, 2005.

[2] C. Clavel, I. Vasilescu, L. Devillers, G. Richard, and T. Ehrette, "Fear-type emotion recognition for future audio-based surveillance systems," *Speech Communication*, vol. 50, no. 6, pp. 487–503, 2008.

[3] G. Valenzise, L. Gerosa, M. Tagliasacchi, and A. Sarti, "Scream and gunshot detection and localization for audio-surveillance systems," in *Proc. IEEE Int'l Conf. Advanced Video and Signal based Surveillance*, 2007.

[4] S. Ntalampiras, I. Potamitis, and N. Fakotakis, "On acoustic surveillance of hazardous situations," in *Proc. IEEE-ICASSP*, 2009.

[5] E. Ohmura, H. Nanjo, H. Kawano, and T. Nishiura, "Fundamental study of automatic gender detection from shout for acoustic-based security system," in *In Proc. WESPAC X*, 2009.

[6] M. Fujimoto and K. Ishizuka, "Noise robust voice activity detection based on switching kalman filter," in *In Proc. INTERSPEECH*, 2007, pp. 2933–2936.

[7] Y. Denda, T. Tanaka, M. Nakayama, T. Nishiura, and Y. Yamashita, "Noise-robust hands-free voice activity detection with adaptive zero crossing detection using talker direction estimation," in *Proc. INTERSPEECH*, 2007, pp. 222–225.

[8] H. Kawano, M. Morise, T. Nishiura, and H. Nanjo, "Fundamental study of radiation characteristics of shouted speech for shouted speech detection towards acoustic-based security system," in *In Proc. WESPAC X*, 2009.

[9] M. Goto, K. Itou, and S. Hayamizu, "Speech completion: On-demand completion assistance using filled pauses for speech input interfaces," in *Proc. ICSLP*, 2002, pp. 1489–1492.

[10] A. Lee, T. Kawahara, and K. Shikano, "Julius – an open source real-time large vocabulary recognition engine," in *Proc. EUROSPEECH*, 2001, pp. 1691–1694.