INTERSPEECH 2009 BRIGHTON

A Fundamental Study of Shouted Speech for Acoustic-Based Security System

Hiroaki NANJO[1], *Hiroki MIKAMI*[1], *Hiroshi KAWANO*[2] *and Takanobu NISHIURA*[2]

[1]Faculty of Science and Technology, Ryukoku University, Japan
[2]Graduate School of Science and Engineering, Ritsumeikan University, Japan

nanjo@rins.ryukoku.ac.jp, nishiura@is.ritsumei.ac.jp

Abstract

A speech processing system for ensuring safety and security, namely, acoustic-based security system 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 based on the understanding of speech events. In this paper, we describe fundamental results of shouted speech.

Index Terms: Security, Shout, Speech understanding

1. Introduction

An acoustic-based security system based on understanding of speech events is addressed. Focusing on indoor security, we study an advanced acoustic-based system that can discriminate emergency shouting from other speech events. In emergency speech, there are two types of speech, shouting and screaming. Shouting contains linguistic information and screaming does not. Conventional studies of shouted speech have mainly focused on speaker identification [1] or event detection [2], and have rarely considered making use of linguistic information included in the shouts. Actually, in such studies, only large sound signals such as screams and gunshots were detected.

In this work, we study how to make use of linguistic information in shouted speech for an advanced acoustic-based security system. Automatic speech recognition (ASR), which converts speech signals to text, is promising for such a security system since the security officers can take appropriate actions according to the ASR results. Although ASR works well for normal speech, it does not work well for excited speech such as shouting by a person crying for help. This paper describes fundamental results for shouted speech characteristics, shout detection, and ASR of shouted speech.

2. Corpus of shouted speech

In this work, isolated shouted words are investigated since people are unlikely to shout sentences in an emergency. Moreover, it is hard to shout sentences in principle. For an actual security system, it is significant to investigate what kinds of words are shouted in an emergency. Here, we constructed a corpus of shouted speech taking account of such a viewpoint. We recorded shouted utterances consisting of 50 Japanese words by 10 male speakers (500 utterances in total). The same 10 speakers uttered the same 50 words normally (500 in total).

First of all, we should investigate and make clear the problems derived from shouting removing other factors. Thus, first, utterances were recorded in an anechoic room. Here, we made three kinds of speech data, which are listed in the upper part of Table 1, so that we could investigate the effects of distance between microphone and speaker. **Set-1** and **set-2** consist of utterances recorded with a headset microphone and a distant microphone (1m), respectively. **Set-3** simulates the utterances that would be recorded with a ceiling-mounted microphone. Specifically, we performed convolution of headset microphone-captured speech with impulse response, which is measured the between the ceiling microphone and the mouth simulator.

Table 1: *Shouting database*

anechoic room	set-1	headset microphone
	set-2	distant microphone (1m)
	set-3	ceiling microphone (set-1+anechoic IR)
real environment	set-4	set-1 + echo room IR (E2A)
	set-5	set-1 + tatami-floored room IR (JR2)
	set-6	set-1 + conference room IR (OFC)

IR: impulse response

Table 2: *Specification of each recording room*

room	reverberation	temperature	humidity	dBA
echo	0.3s	22.1 °C	37.1 %	18.9
tatami-floored	0.47s	12.0 °C	36.8 %	44.8
conference	0.78s	20.7 °C	36.2 %	43.2

We also constructed speech data under real environments for advanced investigations of shouting. Specifically, we performed a convolution of **set-1** speech with impulse responses of three rooms. The details are shown in the lower part of Table 1 (**set-4** to **set-6**). Here, we used impulse responses of the RWCP sound scene database (impulse response and speech data with microphone array) [3]. The specifications of rooms are listed in Table 2.

3. Characteristics of shouted speech

First, we performed a preliminary test with a small shouting database that consists of 10 Japanese words from 9 male speakers recorded with a headset microphone. 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 Table 3. In shouted speech, formants are shifted to the high frequency domain, and standard deviations are larger. Cepstral distance (mean-squared difference) between pairs of vowels is listed in Tables 4 and 5. 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. The power histogram and F0

10.21437/Interspeech.2009-317

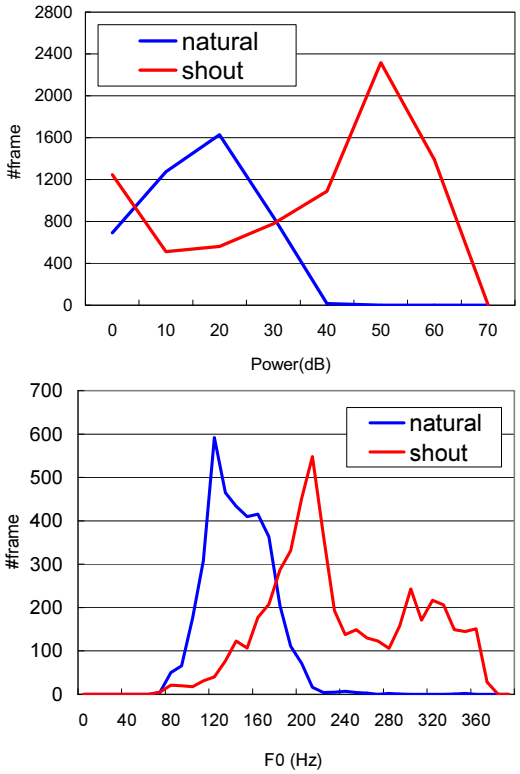

Figure 1: *Power histogram (top) and F0 histogram (bottom)*

histogram are shown in Figure 1. The results also show that shouted speech differs from natural speech, and these features are promising for shout detection.

Table 6 shows the average and variance of cepstral distance for each vowel between speakers. The results suggest that the distributions of vowels are almost the same for shouted and natural speech, and shouted speech recognition could be performed accurately with a speaker-independent acoustic model.

4. Discrimination of shouted speech from natural speech

For security systems, shout detection is significant. Many conventional studies have tried to discriminate voice from other sound signals, namely, voice activity detection (VAD) [4] [5]. In this paper, we do not perform VAD. We assume that VAD could be perfectly performed, and try to discriminate shouting from natural speech. Here, in order to investigate the effects of distance between microphone and speaker, utterances recorded in an anechoic room (**set-1** to **set-3**) are used for the discrimination experiment.

4.1. Discrimination based on detected speech power

As shown in Figure 1, speech power seems to be one of the most significant features for discrimination of shouting and natural speech. However, speech power detected by a microphone is quite sensitive to distance between the microphone and speech source. Figure 2 illustrates the power of shouting detected by a distant microphone (**set-2**) and one of natural speech detected by a headset microphone (**set-1**). It is confirmed that there is a large overlap of both distributions. In an actual security system, we cannot control distances between a microphone and a speaker who is shouting for help. Therefore, discrimination based on detected speech power is unreliable.

Table 3: *Formant of male speakers*

	natural (Hz) F1 / F2	shouted (Hz) F1 / F2
/a/	739.1 / 1376.9	855.4 / 1590.2
/i/	362.9 / 2147.1	427.2 / 2256.3
/u/	351.2 / 1447.3	480.9 / 1524.5
/e/	531.1 / 1952.3	744.5 / 2013.9
/o/	473.9 / 874.0	703.0 / 1193.3

Table 4: *Cepstral distance between pairs of vowels (within-speaker cepstral distance)*

	/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

Table 5: *Cepstral distance between natural and shouted speech (within-speaker cepstral distance)*

/a/-/a/	/i/-/i/	/u/-/u/	/e/-/e/	/o/-/o/
1.99	1.81	1.90	1.48	2.28

Table 6: *Cepstral distance within-vowel (between-speaker cepstral distance)*

		/a/	/i/	/u/	/e/	/o/
natural	mean	1.09	1.03	0.98	1.11	1.19
	var.	0.18	0.09	0.12	0.12	0.18
shout	mean	1.13	0.96	1.23	1.00	1.15
	var.	0.17	0.13	0.20	0.16	0.32

4.2. Discrimination based on spectral envelope

Based on the background, we have investigated the discrimination method without absolute speech power. According to our analyses of shouted speech described in section 3, we adopt an acoustic feature based on the Mel-Frequency Cepstrum Coefficient (MFCC), which models the spectral envelope and is commonly used in ASR. Specifications of speech analysis and acoustic features are described in section 5.1. Using the feature, six acoustic models based on continuous density Gaussian-mixture HMMs were trained with shouting and natural speech of **set-1** to **set-3**. For training, we took the Maximum Likelihood Linear Regression (MLLR) adaptation from the baseline acoustic model described in section 5.1. Here, in order to avoid adapting the acoustic model to the test speaker, 10-fold cross validation is performed. Specifically, for the classification of utterances of a specific speaker, the 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 classification is then performed according to the scores. For example, input speech x is determined as shouting 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 shouting of **set-i** ($i = 1, 2, 3$), and $P(x|\omega^j_{\text{natural}})$ is an acoustic model score derived from HMM trained with natural speech of **set-j** ($j = 1, 2, 3$).

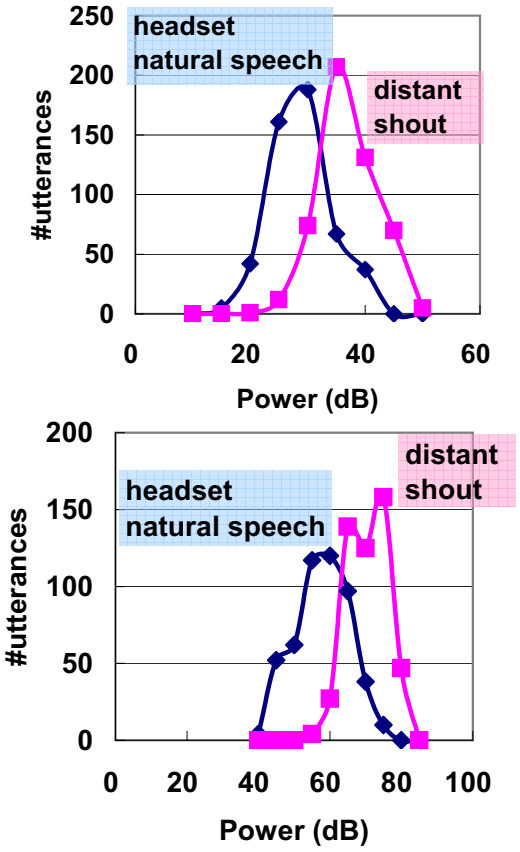

Figure 2: *Distributions of speech power detected by headset and distant microphones (top: averaged power in each utterance, bottom: max. power of each utterance)*

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

Table 7 lists the discrimination results. Here, THRESH is defined empirically for each test set. We confirmed that the false alarm rate is 22.6% to 35.4%, and increases according to the distance between speaker and microphone. This fact shows that we can achieve about a 30% false alarm rate when microphones are placed so that the distance between a person and at least one microphone is less than 1m. Moreover, since almost all speech input to an actually operating security system is expected to be natural speech, the discrimination method can reduce the number of speech events to be checked by a human security officer to about 60 to 80%.

For more accurate detection, we have been investigating the use of other features such as a prolonged vowel [6] in shouting (e.g. "pleeeeeeeease"), radian characteristics [7], speaking rate [8], and spectral movement.

Table 7: *Discrimination of shouting and natural speech*

	natural → natural (correct)	natural → shout (false alarm)
set-1	387 (77.4%)	113 (22.6%)
set-2	333 (66.6%)	167 (33.4%)
set-3	323 (64.6%)	177 (35.4%)

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

Table 8: *Shouted speech recognition result (WAcc. %)*

vocab. size	baseline	MLLR	MAP
50	69.8	94.2	94.2
100	67.0	95.6	94.2
200	63.0	94.6	94.2
300	61.2	93.6	92.8
400	58.8	91.8	93.0
500	60.6	95.0	93.6

test set: set-1

5. Speech recognition for shouting

5.1. Effects of model adaptation and vocabulary size

To understand speech events, ASR is indispensable. In this paper, an isolated word recognition system is set up with three states left-to-right HMM acoustic model and a decoder Julius [9]. As for the acoustic model, a gender independent triphone model (2,000 states, 16 mixtures) trained with the 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 6 lexicons with vocabulary size of 50, 100, 200, 300, 400, and 500. All lexicons include 50 test words, that is, there are no out-of-vocabulary (OOV) words.

ASR is performed for shouted speech of **set-1** using a baseline acoustic model, which is trained with naturally uttered speech. The results are listed in Table 8 ("baseline"). Even for the small vocabulary case, ASR accuracy is 70%, which is not sufficient for a 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 shouting more accurately.

Therefore, we investigated the adaptation of the 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 the acoustic model to the test speaker. The results using an acoustic model adapted to shouted speech by the MLLR and MAP methods are also listed in Table 8. Because of small data size, MLLR outperformed MAP in this experiment, and we achieved about 95% accuracy for both adaptation methods even for a 500-word lexicon. The lexicon size must be sufficient for a security system. The results show that if we capture shouted speech under an ideal condition (headset microphone), we can achieve practical accuracy of shouted speech recognition for a security system.

5.2. Speech recognition under several environments

Next, shouted speech recognition in several environments was investigated. Specifically, we tested with speech of **set-1** to **set-6**. For recognition of each set, an acoustic model adapted to the

Table 9: *ASR results in several environments*

	natural	shouting
set-1	96.0%	94.2%
set-2	97.0%	93.6%
set-3	94.8%	94.2%
set-4	95.6%	90.6%
set-5	93.2%	72.0%
set-6	89.4%	65.4%

Vocabulary size: 50 (no OOV words).
For each set, MLLR is performed (10-fold cross validation)

corresponding test set and a lexicon with a vocabulary size of 50 were used. The results are listed in Table 9.

For speech recorded in an anechoic room (**set-1** to **set-3**), we achieved about 94% ASR accuracy even for shouting. It was confirmed that the distance between speaker and microphone might not affect shouted speech recognition. For speech recorded in real environments (**set-4** to **set-6**), ASR accuracy decreased according to the reverberation time. Note that, for natural speech of **set-5** and **set-6**, we achieved about 90% accuracy. Shouted speech recognition is more sensitive to reverberation time than ASR of natural speech. In a reverberation room, reverberated sound signals of prolonged vowels that have stronger power add to succeeding speech signals, and that is one of the main reasons for ASR degradation for shouting in longer reverberation rooms. Another possible reason is that MLLR adaptation might not work well since shouts recorded in a longer reverberation room are quite different from an acoustic model that was trained with speech recorded in an anechoic room.

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

6. Conclusions

We described the fundamental results of shouted speech for an acoustic-based security system. First, we analyzed the characteristics of shouted speech. Then, we investigated the discrimination of shouted speech from natural speech and the ASR of shouting under several environments. Shouted speech is quite different from natural speech, and we confirmed the significance of further investigations under a real reverberation environment including speech analysis, acoustic modeling unit, and HMM topology.

7. 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, no. 1, pp. 482–486, 2005.

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

[3] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and T. Yamada, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," in *Proc. 2nd Int'l Conference on Language Resources and Evaluation (LREC)*, 2000, pp. 965–968.

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

[5] 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.

[6] 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.

[7] 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 *10th Western Pacific Acoustics Conference (WESPAC X)*, 2009.

[8] H. Nanjo and T. Kawahara, "Language model and speaking rate adaptation for spontaneous presentation speech recognition," *IEEE Trans. Speech & Audio Process.*, vol. 12, no. 4, pp. 391–400, 2004.

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