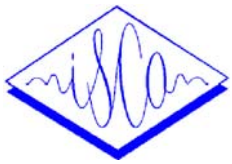

INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

Noise Robust Real World Spoken Dialogue System using GMM Based Rejection of Unintended Inputs

Akinobu Lee, Keisuke Nakamura, Ryuichi Nisimura†, Hiroshi Saruwatari, Kiyohiro Shikano

Graduate School of Information Science, Nara Institute of Science and Technology
Nara 630-0192, Japan
{ri,keisu-na,sawatari,shikano}@is.naist.jp

† Faculty of Systems Engineering, Wakayama University
Wakayama 640-8510, Japan
nisimura@sys.wakayama-u.ac.jp

Abstract

To realize a robust spoken dialogue system for use in a real environment, the robust rejection of *unintended* inputs such as laughter, coughing, background speech and other noise based on GMM is implemented and examined on the basis of actual utterances. All the triggered inputs to a speech-oriented guidance system from 125 days of field tests in a public space are collected, and the occurrence of unintended inputs is investigated. GMM classifiers for voice categories (adult speech and child speech) and non-voice categories (laughter, coughing and other noises) are trained on the basis of the analysis result. The rejection performance of unintended speech was experimented on actual uncontrolled real inputs, and an EER of 3.32% was achieved by the 5-class GMM, which outperforms simple 2-class (voice / non-voice) GMM. The rejection of background speech using GMM is also investigated.

1. Introduction

To realize a natural speech interface in our daily environment, a speech interface system should be able to deal with a large number of mis-triggered, non-speech inputs such as impulsive noise, background speech, laughter, coughing, and many other sounds that inevitably intrude into the microphone. In an uncontrolled environment, such inputs should be discriminated from a user's speech inputs to render the speech interface more robust and reliable. In this paper, speech inputs which reflects a user's intention of speaking to the system is called *intended* speech, regardless of the task domain, and other accidental, unintentional, or wrongly triggered inputs such as noise, laughter, coughing and background speech are called *unintended* input.

Rejection of invalid inputs has been a major concern in the development of a practical speech recognition system, especially as a problem of utterance verification. The inputs are typically examined using acoustic confidence measures [1] which is derived from a recognizer, or utterance-level likelihoods for out-of-domain utterance rejection [2]. Another approach may be from multimodal schemes using face direction detection or lip image recognition, but their application is generally limited to controlled environments and requires extra equipment.

In this work, a verification of intended input based on Gaussian Mixture Model (GMM) in an actual natural spoken dialogue system is investigated. We are developing a speech-oriented information guidance agent for a public space (the civil hall of Ikoma city, Nara, Japan). It has run throughout business hours since November of 2002, without any operator, and has collected a large number of natural human-to-machine speech interactions and has also recorded various types of inputs [3]. This work is motivated by the progress in the field of speaker verification using GMM [4], since GMM has been proved to be a powerful tools for a text-independent speaker verification.

This paper consists of five major sections. The second section describes the specifications of our spoken dialogue system and its location, and the third section presents the details of the collected data. The fourth section gives the specifications of the GMM based rejection of unintended input. The experimental results are presented in the fifth section, and we conclude our work in the final section.

2. Speech oriented information guidance system "Takemaru-kun"

In order to collect data and conduct a field test to determine how input occurs when users face a real spoken dialogue system in a natural environment, a speech-oriented information guidance system for public use called "Takemaru-kun" has been developed [3]. It has been located during business hours at the entrance hall of the Ikoma Community Center (Figure 1) since November 2002. It can inform visitors about the center and Ikoma city via a speech interface, with animated characters and related WWW pages. It implements a simple one question and one answer strategy, and can respond to questions about the facilities of the hall, information about shopping around the hall, traffic information and so on. It can also respond to greetings and personal questions about the agent Takemaru-kun itself. A user's speech is captured by a single microphone, and the response is output by a synthesized speech generated by Text-To-Speech software and animated gestures provided along with related WWW pages on the screen. Figure 1 shows a snapshot of Takemaru-kun for the setup in the community center.

The speech recognizer is our open-source speech recognition engine Julius [5]. For the language model, a task-dependent word 3-gram is trained by Web texts related to Ikoma city and transcriptions of typical questions for the task, collected

Figure 1: Speech-oriented information guidance agent "Takemaru-kun."

by hand. The 3-gram is further adapted using task grammar. The dictionary size is about 40,000 words. For the acoustic model, a speaker independent phonetic tied-mixture (PTM) triphone model is used, which is trained by the JNAS newspaper database with a 25dB SNR exhibition hall noise superimposed over the training set. See [3] and [6] for detailed specifications.

3. Data collection and analysis

The system has been located at the entrance hall of the Ikoma Community Center, and all the input segments triggered by the recognizer were recorded. The input detection method is based on a raw level threshold and zero-cross count. The system has been in operation at the entrance hall daily during business hours, with no human operator on the side [6]. In this study, data from the first four months, from November 6, 2002 to March 21, 2003, are investigated. A total of 46,755 inputs were recorded, which had an average of 374 utterances per day.

The collected inputs are classified manually to examine how unintended inputs occur. Noise inputs that contains no speech part are classified into the "Noise" category. Otherwise, if the utterance is clear and can be easily transcribed by a human, and its speaker has the obvious intention of talking to the agent, it is classified into the "Clean utterance" category. Fillers and disfluencies also belong to this category. Other inputs should be further classified into the "Wrongly triggered utterance" and "Non-verbal" catebories. The former should contain unintentionally triggered utterances (i.e., background speech) and incomprehensible speech (level underflow, level overflow and vocal sounds which are impossible to transcribe). The latter consists of "laughter" and "coughing". The utterances in the "Clean utterance" category are further classified into "adult" and "child" according to their speech characteristics. If an input belongs to several categories above, i.e. a speech that contains both laughter and clean utterance, the most dominant characteristics are chosen. The classification of all inputs was carried out by one operator, by hand.

The classification result is summarized in Table 1. It was found that clean, valid utterances were only 69.1%, with the remainder consisting of unintended, invalid inputs that should be rejected. In particular, background speech and laughter occur very often. Although the users are not conscious of making any inputs to the system, these unintended, mis-triggered inputs may confuse the recognition system and cause it to respond incorrectly, which often results in the user's irritation, annoyance and uncertainness toward a speech interface. In order to realize a practical speech recognition system that is robust, reliable and easily accessible to everyone, the system should have the ability to detect and reject unintended and invalid inputs.

Table 1: Classification of Collected data

Class of input	# of inputs
Intended	
Clean utterance	
adult	8,672
child	23,638
Unintended	
Wrongly triggered utterance	
background speech	4,929
level underflow	1,469
level overflow	50
incomprehensible	309
Non-verbal	
laughter	1,091
coughing	198
Noise	6,399
Total	46,755

4. GMM-based unintended input rejection

Upon classification of the collected data, a GMM-based speech verification to detect clear and intended inputs is investigated. The Gaussian Mixture Model (GMM) has been a major tool for noise / speech verification and, therefore, for speaker identification, because of its text-independency and powerful discrimination performance. Thus, it is also promising to use GMM to discriminate the acoustic properties of the invalid inputs. Conventional studies in speech verification have focused on the rejection of environmental noise such as impulsive noise, music and applause. In this paper, the main focus is on the rejection of more utterance-like wrong inputs such as laughter, coughing and wrongly triggered background speech. Our aim is to be able to perform text-independent verification of clean utterances and reject invalid inputs using GMM.

GMMs were trained according to the the classification in Table 1. The training set was defined according to the amount of data in each category. Data that were not strictly determinable were excluded from the training, and classes with small amount of data are gathered in order to obtain sufficient amount of training data. As a result, five classes are defined: child, adult, laughter, coughing and noise. Table 2 shows the training conditions and the amount of data set. "Intended / Valid" is a category of to-be-recognized input, and "Unintended / Invalid" is for inputs that should be rejected before the recognition process.

GMMs with mixtures of 2 to 512 are built in order to see the trade-off between discrimination ability and processing cost. For comparison, two GMMs corresponding to the "Intended / Valid" and "Unintended / Invalid" categories are also trained by using all the data in their sub classes. For the laughter and coughing data, samples that contain overlapping speech are excluded from training.

Table 2: Training condition of GMM

Amount of Training data	Intended / Valid	child	20,016
		adult	4,065
	Unintended / Invalid	laughter	849
		coughing	98
		noise	6,413
Sampling rate/bit	16kHz, 16bit		
Window width/shift	25/10 msec		
Parameter	MFCC (12 dim.), ΔMFCC, Δpower		

"other" includes background speech and noise.

Table 3: Test set specification

Class	child	adult	laughter	coughing	noise
# of input	500	500	100	100	100

5. Experiments

Speech verification and unintended speech rejection based on the GMM were carried out experimentally through identification and verification tests. Test set samples are extracted from the collected data, and these were excluded from the training set. Details of the test set are shown in Table3. The class names of the test set are the same as those for the training data.

5.1. 5-class identification

First, 5-class identification was examined in order to observe the discrimination performance of the GMMs. The 5 GMMs trained as described in section 4 were applied to each input, and the one with the best acoustic likelihood was taken as giving the identification result. When applying the child, adult and laughter GMMs, the head and tail silences in each input are matched by silence GMMs, "silB" and "silE", respectively, which were trained by using the head and tail silences in the training data.

Figure 2 shows the identification performance for various numbers of Gaussian mixtures. Given a sufficient number of mixture components, an error rate of less than 15% was achieved for all of the classes except coughing. It is also found that laughter can be discriminated from normal speech with a high degree of accuracy, almost the same as that of normal speech. The performance of adult and child GMMs was saturated for small mixture components, whereas the laughter and noise GMMs require many mixture components to achieve a sufficiently high discrimination ability, since the noise and laughter categories cover a wide variety of sounds. The insufficient performance of the coughing GMM may result from its small number of training data.

5.2. Unintended input rejection

Next, the rejection of unintended inputs was carried out. The verification of a given input, determination whether it should be accepted as intended speech or rejected as an unintended input, was carried out by comparing the best likelihood of valid GMMs (adult and child) with the best likelihood of invalid GMMs (laughter, coughing and noise).

The equal error rate (EER) using the class-level 5 GMMs is plotted in Figure 3. Performance using a category-level 2 GMMs (valid / invalid) is also plotted for comparison. In all mixtures, the 5-class method outperforms the 2-class method.

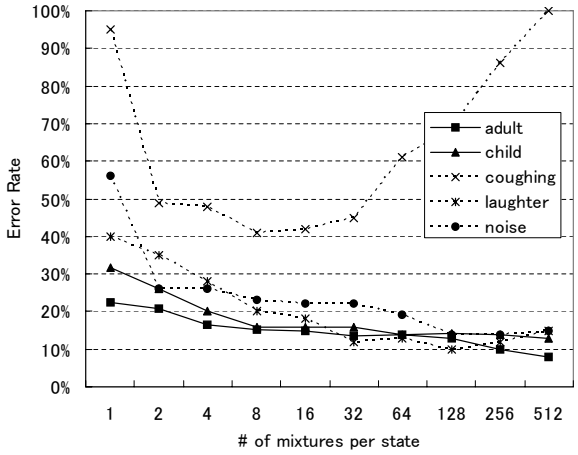

Figure 2: 5-class identification result for various mixtures of GMM.

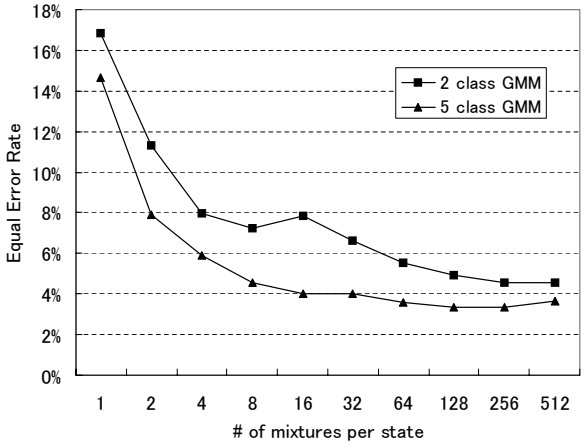

Figure 3: Performance of speech verification.

The optimal EER of the 2-class method was 4.53%, whereas the 5-class method achieved 3.32%, which is an improvement in relative error of 27%. The superiority of 5-class method was remarkable for the laughter test set, since the acoustic properties of laughter are more similar to speech than to noises, and almost all of the laughter test set was accepted as valid speech inputs in the 2-class method. Also, Figure 4 shows the detection error trade-off (DET) curves [7] for several conditions. It is evident that the 5-class 16-mixture GMM has almost the same performance as the 2-class 128-mixture GMM, and the 5-class 128-mixture GMM showed the best result.

The detailed results of the 5-class 128-mixture GMM are shown as a confusion matrix in Table 4. Most of the incorrectly identified coughing samples were captured by both the laughter and noise models, which resulted in a successful rejection. This result suggests that custom design of the category and class of the GMM for the target environment can be a primary factor in achieving successful GMM-based robust verification.

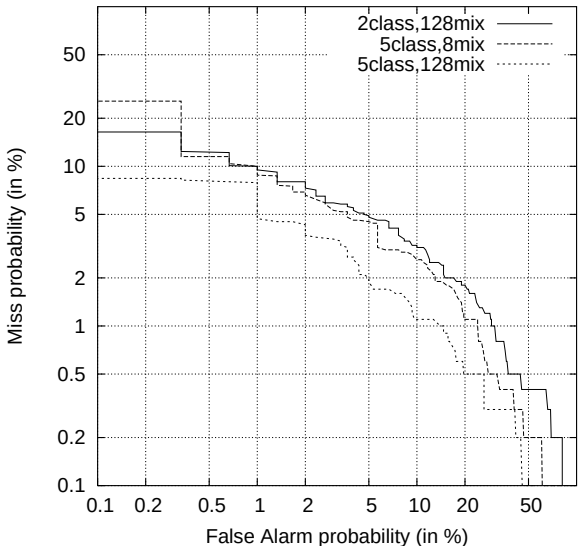

Figure 4: DET curves.

Table 4: Confusion matrix

	Result				
	Valid		Invalid		
test set	adult	child	laughter	coughing	noise
adult	436	58	2	0	4
child	51	429	11	0	9
laughter	0	2	90	0	8
coughing	1	2	33	31	33
noise	3	4	7	0	86

5-class GMM, 128 mixtures

5.3. About background speech

The identification and rejection of background speech is further investigated. Here we classified the background speech inputs in the test set into three sub-categories, according to there clarity: (bg-1) far distant speech which is difficult to transcribe, (bg-2) near distant speech with some difficulty in transcribing, and (bg-3) close talk, which is clean but unintended speech. Figure 5 shows the score distributions of samples in each sub-category for the noise GMM. For reference, the distributions of noise and voice (adult and child) inputs are also plotted. It is suggested that the bg-1 samples are almost as discriminable as the noise inputs, and the bg-2 samples can be discriminated if we build a specific GMM for that purpose. It is also confirmed that the clean bg-3 samples cannot be detected by GMM, and therefore they should be identified on the basis of their linguistic information, just like an out-of-domain utterance verification using confidence measures.

6. Conclusions

GMM-based robust speech verification for a real-world spoken dialogue system was developed, and its ability to discriminate and reject unintended inputs was investigated through application to actual natural human-to-machine utterances. In order to accept only clear intended utterances and reject non-verbal and wrongly triggered inputs, the GMMs were trained for adult speech, child speech, laughter, coughing and noises, according to the classification of he field test data. An equal error rate of 3.32% was achieved using the 5 GMMs, which outperform simple two-class (valid / invalid) GMMs. In particular, it was found that laughter can be discriminated with the GMMs with a high degree of accuracy. Rejection of background speech is also possible in this scheme, although an integration with another out-of-domain utterance verification scheme is considered necessary.

Future work will be dedicated to the automatic clustering of a training set to obtain an optimal GMM set, and the utilization of detecting non-verbal inputs such as laughter, toward a more natural and reliable spoken dialogue interface.

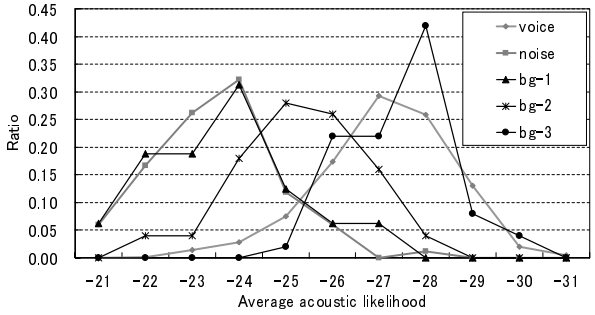

Figure 5: Acoustic likelihoods of background speech on noise GMM.

7. Acknowledgements

Part of this work is supported by MEXT e-Society leading project.

8. References

[1] Z. Rivlin, M. Cohen, V. Abrash and T. Chung. A phone-dependent confidence measure for utterance rejection. in Proc. IEEE-ICASSP, pages 515–517, 1996.

[2] M. G. Rahim, C.-L. Lee and B.-H. Juang. Discriminative utterance verification for connected digits recognition. in IEEE Trans. on Speech and Audio Processing, vol.5, no. 3, pages 266–277, 1997.

[3] R. Nisimura, Y. Nishihara, R. Tsurumi, A. Lee, H. Saruwatari, and K. Shikano. Takemaru-kun: Speech-oriented information system for real world research platform. in *Int'l Workshop on Language Understanding and Agents for Real World Interaction*, pages 70–78, 2003.

[4] D. A. Reynolds. Speaker identification and verification using Gaussian mixture speaker models. in *Speech Communication*, vol.17, pages 91–108, 1995.

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

[6] R. Nisimura, A. Lee, H. Saruwatari, and K. Shikano. Public speech-oriented guidance system with adult and child discrimination capability. in *Proc. IEEE-ICASSP*, pages 433–436, 2004.

[7] A. Martin, G. Doddington, T. Kamm, M. Ordowski and M. Przybocki. The DET curve in assessment of detection task performance. in Proc. EUROSPEECH, pages 1895–1898, 1997.