INTERSPEECH 2009 BRIGHTON

System Request Detection in Human Conversation Based on Multi-Resolution Gabor Wavelet Features

Tomoyuki Yamagata, Tetsuya Takiguchi, Yasuo Ariki

Department of Computer Science and Systems Engineering, Kobe University
1-1 Rokkodai, Nada, Kobe, 657-8501, Japan
yamagata@me.cs.scitec.kobe-u.ac.jp, takigu@kobe-u.ac.jp, ariki@kobe-u.ac.jp

Abstract

For a hands-free speech interface, it is important to detect commands in spontaneous utterances. Usual voice activity detection systems can only distinguish speech frames from non-speech frames, but they cannot discriminate whether the detected speech section is a command for a system or not. In this paper, in order to analyze the difference between system requests and spontaneous utterances, we focus on fluctuations in a long period, such as prosodic articulation, and fluctuations in a short period, such as phoneme articulation. The use of multi-resolution analysis using Gabor wavelet on a Log-scale Mel-frequency Filter-bank clarifies the different characteristics of system commands and spontaneous utterances. Experiments using our robot dialog corpus show that the accuracy of the proposed method is 92.6% in F-measure, while the conventional power and prosody-based method is just 66.7%.

Index Terms: dialog system, voice activity detection, system request detection

1. Introduction

Recently, speech interfaces are usually applied to devices that users cannot operate by hands, such as car navigation systems and robots. These systems usually utilize a voice detection system in order to discriminate human speech from background noises, e.g. [1, 2, 3, 4, 5]. However, it may be difficult for these interfaces to discriminate system requests - utterances that users speak to the system - from human-human conversations. Therefore, a current speech interfaces for car navigation systems require a physical button that needs to be pushed in order to switch the microphone input on and off. If there is no such button, all conversations are recognized as commands for the system. The need to push a button, however, eliminates the merit of so-called hands-free speech interfaces since users still need to operate the system, to some extent, by hand.

Speech Spotter [6] is one solution to the problem. However, Speech Spotter requires users to consciously change their speaking style. Related to this issue, research has been carried out on ways of discriminating system requests from human-human conversations using acoustic and prosodic features calculated from each utterance [7]. There are also discrimination techniques using speech recognition-based linguistic features. Keyword or key-phrase spotting-based methods [8, 9] have also been proposed. However, when utilizing the keyword spotting-based method, it is difficult to distinguish system requests from explanations about how to use the system - utterances made by users when explaining the system to another person. It becomes a problem when both types of utterances contain the same "keywords". For example, the request speech is "Come here", and the explanation speech is "If you say that come here, the robot will come here." In addition, there are various costs involved with constructing a network grammar to accept flexible expressions.

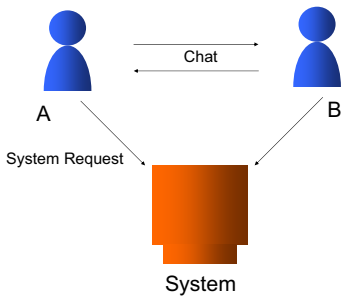

Figure 1: Two-person-and-one-system dialog

2. Recording Conditions and Details of Corpus

In our previous work [10], we proposed an acoustic-based feature for discriminating commands from human-human conversations, where the head and the tail of an utterance are considered. However, as only simple power and pitch features were used as acoustic features, the performance was not adequate. In this paper, in order to analyze the difference between system requests and spontaneous utterances, we focus on fluctuations in a long period, such as prosodic articulation, and fluctuations in a short period, such as phoneme articulation.

Gabor features have variable time and frequency resolution

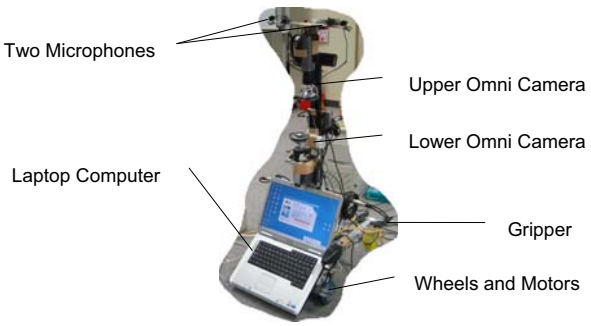

Figure 2: Photo of mobile robot

Table 1: Function list of the mobile robot

Functions	Sound source direction presumption based on CSP
	Move toward/backward sound source
	Obstacle avoidance
	Put down a bottle with the gripper
	Take a face picture
Command examples	"*Kotchi ni kite.*" (Come here.)
	"*Mukou e itte.*" (Go to the other side.)
	"*Shashin wo totte.*" (Take my picture.)
	"*Watashi ni tsuite kite.*" (Come with me.)
	"*Bottle wo oite.*" (Put down the bottle.)

Table 2: The number of utterances and system requests

Total utterances	System requests
1,024	110

Table 3: Prosodic features used in a baseline

Power	Ave.	S.D.	Max.	Max. − Min.
Pitch	Ave.	S.D.	Max.	Max. − Min.

Ave.: Average
S.D.: Standard deviation

and have been used as speech features in speech recognition systems [11, 12, 13]. In this paper, we describe an advanced method of discrimination using only acoustic features based on multi-resolution analysis using Gabor wavelet on a Log-scale Mel-frequency Filter-bank. The detailed acoustic and prosodic analysis improve the system request detection accuracy significantly.

The corpus for evaluation was recorded under these conditions: two people (speakers) and a system are located in the same place, as shown in Figure 1. The two people talk to each other and sometimes make requests to the system. This situation is quite common: for example, two people talking in a car and also operating a speech-activated car navigation. In this paper, we used a mobile robot as the system, because recording in a real car causes noise problems. Our task was to detect system requests from among various spontaneous human utterances.

A photo of the robot is shown in Figure 2. It is equipped with two microphones (different from recording microphones), two omni cameras (upper view and lower view), a laptop computer to control the system, a gripper to hold and put a bottle down, wheels and motors (advancement, retreat, rotation). The functions of the robot are summarized in Table 1. In general, the robot is operated by speaking commands a few meters away from it.

The recording microphones were attached to the chest of each speaker. Three 30-minute recording sessions were held, where utterances were recognized using an automatic speech recognition system, and the mobile robot was also working. The

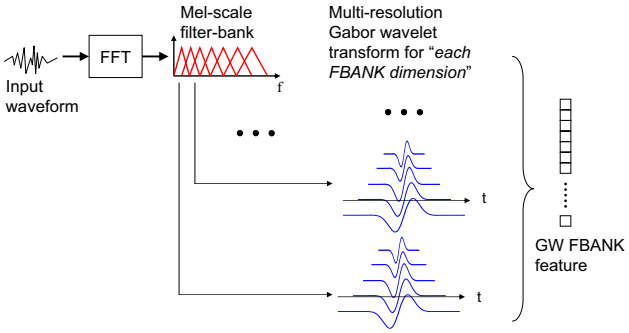

Figure 3: Multi-resolution feature extraction for each FBANK dimension

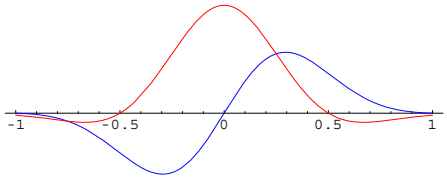

Figure 4: Gabor mother wavelet

total number of speakers was two. We did not show them the list of commands that the robot can accept. One reason we did it this way is to increase the variation of system request commands. The other reason is that we are going to develop speech interfaces that accept not only specific commands but also various expressions. Therefore, the speakers could give commands that might be acceptable to the robot. We labeled those utterances as system requests manually.

Table 2 shows the results of cutting out utterances from the record using power and zero-crossing. The experiments in this paper are performed using 10-fold cross-validation.

3. Previous Feature Extraction Method

In previous research, it was found that the difference between commands and human-human conversations appears in the power and the pitch of the utterance [7]. Therefore, we detected utterance sections using power and zero-crossing, and then calculated the 8 dimensional features shown in Table 3 for each utterance as the baseline. The power was computed by Root Mean Square (RMS). The pitch was calculated by LPC residual correlation.

4. Multi-Resolution Features

The prosodic features described in Section 3 are obtained from raw waveforms. However, it is difficult to say that these features are clearly focusing on speech components. In order to extract the difference between commands and human-human conversations more precisely, we propose a multi-resolution analysis using Gabor wavelet transform based on a Log-scale Mel-frequency Filter-bank (FBANK). In our method, a one-dimensional wavelet transform is applied to the FBANK feature in order to analyze the utterance-based acoustic fluctuation.

Figure 3 shows the flow of our feature extraction. For each dimension of FBANK features, we perform a multi-resolution one-dimensional Gabor wavelet transform. The equation for the

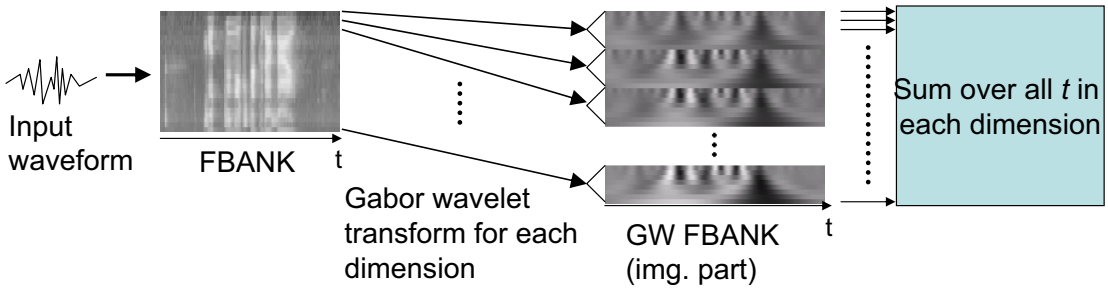

Figure 5: Flowchart of calculating multi-resolution features

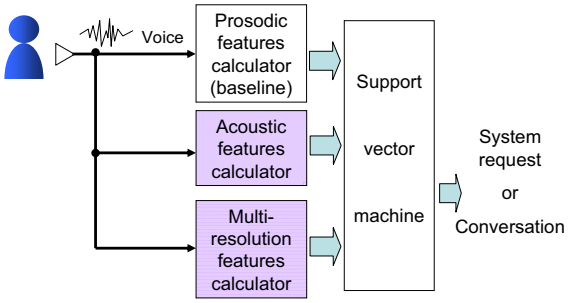

Figure 6: System overview

Table 4: Experiment results of 10-fold cross-validation for system request discrimination

	Precision	Recall	F-measure
Baseline	0.584	0.806	0.667
Filter-bank	0.719	0.909	0.803
Filter-bank + Delta	0.864	0.882	0.874
Filter-bank + Delta (Multi-Resolution)	0.909	0.909	0.909
Gabor Wavelet Re.	0.906	0.873	0.889
Gabor Wavelet Img.	0.933	0.891	0.912
Gabor Wavelet Re. + Img.	0.943	0.909	0.926

Gabor mother wavelet used in this paper is shown in below:

$$\Phi(t) = \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{t^2}{\sigma^2}) exp(j2\pi ft) \quad (1)$$

and the waveform is shown in Figure 4. The red line shows the real part of Gabor wavelet, and the blue line shows the imaginary part. As the imaginary part is an odd function, the transform works like differential calculus. In this paper, the real part and the imaginary part are used independently as multi-resolution features.

In order to deal with fluctuation of each FBANK dimension in various length windows, the σ Gaussian window length is changed for the 8 levels from 512 to 45 [ms] scaled by $1/\sqrt{2}$, and the $1/f$ period is also changed for the 8 levels from 1,024 to 90 [ms] by $1/\sqrt{2}$. A sample transform result is shown in Figure 5 (imaginary part). The coefficients for the real part and imaginary part are independently summed up for all t in each dimension, and those are used as multi-resolution features in this paper.

5. Experiments

The overview of our system request detection system is shown in Figure 6. The experiments are performed for each baseline and proposed feature in order to compare the accuracy of the features. The baseline features are obtained from the input waveform using the method described in Section 3. The 24-dimension FBANK features and the energy are obtained from the "Acoustic features calculator". Since the total number of acoustic feature (FBANK) dimensions is 25, the total number of dimensions of multi-resolution features is 200 (25 $\times$ 8 levels). The features obtained from each utterance are classified by Support Vector Machines. We used SVM^{light} for the Support Vector Machine with the RBF (Gaussian) kernel.

Table 4 and Figure 7 show the experiment results of 10-fold cross-validation for the corpus as shown in Table 2. The experiment results show the cases in which the F-measure - the harmonic average of precision and recall obtained from the equation as below - became the maximum value.

$$F = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (2)$$

As shown in Figure 7, while the accuracy of conventional prosodic features (baseline) is 66.7 in F-measure, the acoustic features on the Log-scale Mel-frequency Filter-bank (25-dimension FBANK) are 80.3. The accuracy of the acoustic features itself is quite high; however, performing multi-resolution Gabor wavelet analysis (Gabor) improves the accuracy to 92.6 without any features combination.

In Figure 7, we compare the result to our previous work [10]. In [10], we focus on the different characteristics of commands and human-human conversations which usually appear on the head and the tail of the utterance, and the prosody features are calculated from three sections (the head, tail, and middle sections of the utterance). Also, our previous paper has described that considering the alternation of speakers using two channel microphones (turn-taking parameters) improved the performance. As shown in Figure 7, compared with the result of our previous work [10], our new method improves the accuracy from 85.1 to 92.6.

In Figure 7, we also compare the results for static FBANK features to delta FBANK features and multi-resolution delta performance. The delta FBANK uses ± 3 frames. The multi-

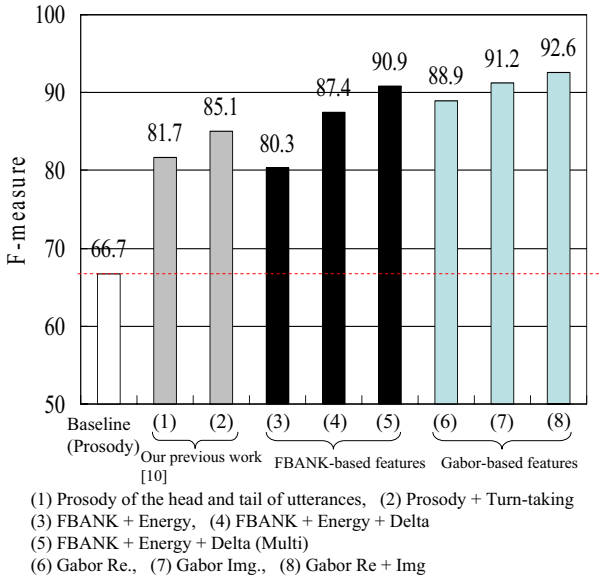

Figure 7: Results of utterance verification

resolution window length delta is set to $\pm$ 3, 4, 6, 8, 12, 16, 23, and 32 frames, where the window length increases from 2 by a factor of $\sqrt{2}$, and these conditions are the same (increasing scale and the 8 levels) as those of the Gabor analysis. The multi-resolution delta provides better performance for the FBANK features.

We also show the results obtained using only the real part (Gabor Re.) or imaginary part (Gabor Img.) of multi-resolution features. These results show that considering only the imaginary part is better than using only the real part, and it suggests that the fluctuation of FBANK features plays an important role in the detection of system requests.

6. Conclusions

In this paper, we describe a multi-resolution-based feature for detecting system requests in an environment that also contains human conversation. To discriminate commands from human-human conversations more efficiently than when using conventional acoustic and prosodic features, it is necessary to consider the variable time and frequency resolution of an utterance.

Using Log-scale Mel-frequency Filter-bank (FBANK) features improves the performance. Because FBANK features are adjusted to capture speech components precisely, the power of utterances calculated from each frequency band extracts extracting speech components more accurately. Also, analyzing FBANK features with multi-resolution Gabor wavelet transform improves the performance even more, where we focus on fluctuations in a long period, such as prosodic articulation, and fluctuations in a short period, such as phoneme articulation in order to analyze the different characteristics of commands and human-human conversations.

Future work will include evaluation under conditions in which the system accepts many kinds of commands under noisy conditions. The improvement of detecting utterance sections and combining linguistic features or other features (e.g. [14, 15, 16]) are also themes that will need to be researched.

7. References

[1] J. Ramirez, J. C. Segura, C. Benitez, A. Dela Torre, and A. Rubio, “An effective subband OSF-based VAD with noise reduction for robust speech recognition,” IEEE Trans. on Speech and Audio Processing, Vol. 13, Issue 6, pp. 1119-1129, 2005.

[2] D. Cournapeau and T. Kawahara, “Using variational Bayes free energy for unsupervised voice activity detection,” Proceedings of ICASSP, pp. 4429-4432, 2008.

[3] N. Kitaoka, T. Yamada, S. Tsuge, C. Miyajima, T. Nishiura, M. Nakayama, Y. Denda, M. Fujimoto, K. Yamamoto, T. Takiguchi, S. Tamura, S. Kuroiwa, K. Takeda, and S. Nakamura, “Development of VAD Evaluation Framework CENSREC-1-C and Investigation of Relationship Between VAD and Speech Recognition Performance,” Proceedings of ASRU, pp. 607-612, 2007.

[4] M. Y. Choi, H. J. Song, and H. S. Kim, “Speech/Music Discrimination for Robust Speech Recognition in Robots,” Proceedings of IEEE International Symposium on Robot and Human interactive Communication, pp. 118-121, 2007.

[5] H. Sakai, T. Cincarek, H. Kawanami, H. Saruwatari, K. Shikano, and A. Lee, “Voice activity detection applied to hands-free spoken dialogue robot based on decoding using acoustic and language model,” Proceedings of International Conference on Robot Communication and Coordination, 2007.

[6] M. Goto, K. Kitayama, K. Itou, and T. Kobayashi, “Speech Spotter: On-demand Speech Recognition in Human-Human Conversation on the Telephone or in Face-to-Face Situations,” Proceedings of ICSLP, pp. 1533-1536, 2004.

[7] S. Yamada, T. Itoh, and K. Araki, “Linguistic and Acoustic Features Depending on Different Situations - The Experiments Considering Speech Recognition Rate,” Proceedings of Interspeech, pp. 3393-3396, 2005.

[8] T. Kawahara, K. Ishizuka, S. Doshita, and C.-H. Lee, “Speaking-style Dependent Lexicalized Filler Model for Key-phrase Detection and Verification,” Proceedings of ICSLP, pp. 3253-3259, 1998.

[9] P. Jeanrenaud, M. Siu, J. R. Rohlicek, M. Meteer, and H. Gish, “Spotting events in continuous speech,” Proceedings of ICASSP, Vol. 1, pp. 381-384, 1994.

[10] T. Yamagata, A. Sako, T. Takiguchi and Y. Ariki, “System Request Detection in Conversation Based on Acoustic and Speaker Alternation Features,” Proceedings of Interspeech, pp. 2789-2792, 2007.

[11] J. N. Gowdy and Z. Tufekci, “Mel-Scaled Discrete Wavelet Coefficients for Speech Recognition,” Proceedings of ICASSP, pp. 1351-1354, 2000.

[12] S. Y. Zhao and N. Morgan, “Multi-Stream Spectro-Temporal Features for Robust Speech Recognition,” Proceedings of Interspeech, pp. 898-901, 2008.

[13] B. T. Meyer and B. Kollmeier, “Optimization and Evaluation of Gabor feature sets for ASR,” Proceedings of Interspeech, pp. 906-909, 2008.

[14] T. Takiguchi, A. Sako, T. Yamagata, and Y. Ariki, “System Request Utterance Detection Based on Acoustic and Linguistic Features ,” Chapter on Speech Recognition, Technologies and Applications. Book edited by F. Mihelic and J. Zibert. pp. 539-550, 2008.

[15] Y. Obuchi, M. Togami, and T. Sumiyoshi, “Intentional Voice Command Detection for Completely Hands-Free Speech Interface in Home Environments,” Proceedings of Interspeech, pp. 119-122, 2008.

[16] M. G. Rahim, C.-H. Lee, B.-H. Juang, “Discriminative utterance verification using minimum stringverification error (MSVE) training,” Proceedings of ICASSP, pp. 3585-3588, 1996.