A REAL TIME LANGUAGE AND DIALECT IDENTIFICATION BASED ON THE VQ ERROR OF THE ACOUSTIC FEATURE AND PROSODIC CUES

I.Dawa , K.Shirai

Department of Information and Computer Science, Waseda University

3-4-1 Okubo, Shinjuku-Ku, Tokyo, 169-8555 Japan

dawa@shirai.info.waseda.ac.jp

ABSTRACT

This paper proposed an application for the language and dialect recognition in spontaneous speech on the real time. The language will be identified based on the VQ (Vector Quantization) error and the fundamental frequency. And the dialect recognized based on the VQ error and the sample different in speaking rate. The reported application system in this paper performed on the seven languages in Asia: Chinese, Japanese, Korean, Thai, Mongolian, Uigur and Khazak; And the dialect performed on the four dialects of Mongolian, such as Kharha of the Mongolia, Chahar of the Inner Mongolian, Oirat of the Xin Jiang in China, and Kalmyk of Kalmykia in Russia. Results from a number of experiments showed that the misidentification rate for language is less than 3% in less than 3 seconds utterance.

1.INTRODUCTION

Development of a multilingual or multidialectal translation system is one of the ultimate goal of researches in natural language processing, speech recognition, and artificial intelligence. It would be preferable that there is an automatic identification faculty for the input language or dialect before speech recognition. It is particularly suitable for Mongolian, which has several dialectal variations in its linguistic, phonetic and writing system, used in some countries and regions[1]. All of these technologies need high accuracy and real-time performance. Here we report an approach for a real-time identification for the multilingual speech input based on the VQ error of acoustic feature and prosodic cues of utterance. This system will expect to be widely applicable in some fields, such as language or dialect identification, spontaneous speech speaker recognition, and segmentation of the long time speech materials uttered by limited speakers in the TV or Radio broadcasting news[2].

2.APPROACHES

The system reported in this paper works in two steps, Firstly, the language is recognized by the inputted utterance, then do the dialect if it need.

2.1 The preceding process

Some of codebooks C_i^L (called centroid vector), which is made by clustering the acoustic future, such as mel-cepstrum, delta mel-cepstrum using about 30 seconds speech sample different in speaking rates, and the fundamental frequencies f_0^L (i *=1,…,N size of codebook vectors, L=1,…k number of languages*), are taken from the speech data given in languages and dialects.

2.2 Calculation of VQ error

The errors between the centroid vector C_i^L, and the test vector x$_j$(j=1,…,M) which is made by acoustic parameter inputted utterance is calculated by the following formula (1) and saved in order of min to max.

$$D_L = \frac{1}{M}\sum_{j=1}^{M} \min_{1\le i\le N}\left[d\left(C_i^L, x_j\right)\right] \quad (1)$$

ISBN 1-876346-48-5 © 2003 UAB

2.3 Decision accept and reject

Fig.1 shows the min VQ errors of neighboring that the system is working normally or the language was recognized correctly on utterances of various speaking rates. Speaking rate was defined as the number of morae spoken per second, and computed for all sentences. We know here, there is an established distortion or error δ_1 between the average_2 and average_1. it will be set as the first threshold in the system.

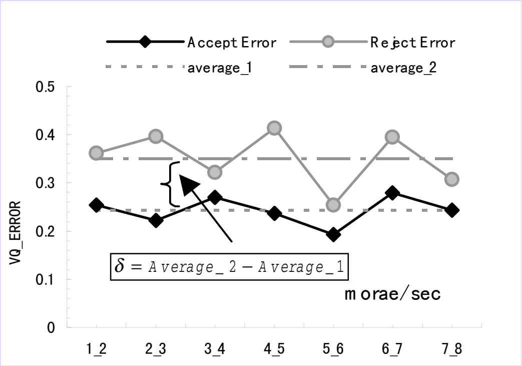

Fig.1 Minimum VQ error in D_L

For the one of inputted utterance, firstly, the min value of the errors $D_{\min}^{k}$ and the neighboring one $D_{\min}^{k+1}$ are selected. Next, if the difference $D_{\min}^{k+1} - D_{\min}^{k} < \delta_1$ is true, then, the language on the $D_{\min}^{k}$ is accepted as the most likely candidate. Else, the decision by prosodic cues is performed for the language of corresponding $D_{\min}^{k}$ and $D_{\min}^{k+1}$, that is, firstly the mean difference,

$$f_0^x = \frac{1}{M}\sum_{j=1}^{M}(|f_o^{k,k+1} - f_0^x|) \quad (2)$$

is calculated, and then the language corresponding the min value in f_0^x is selected as the last candidate.

2.3 Distinction dialects for Mongolian

Dialect differences that exist among ethnic groups can be characterized by the distinction in lexicon, phonology, morphology and prosody[3]. Looking specifically at Mongolian dialects, it is able to detect some distinction between Mongolian dialects in their speaking rate of speech[4]. In case of Mongolian dialects recognition, we found that the recognition error was occurred frequently at the high speaking rate of utterance. The Fig.2 is a real test for Mongolian three main dialects such as Kharha of Mongolia (A_1), Chahar (B_1) and Oiart (C_1) of Inner Mongolian and Xin Jiang in China. It is clear in Fig.2 that VQ error was normal and the dialects were recognized correctly in areas of the low speaking rate (about 1-6 morae/sec), however, VQ errors were increased greatly and the recognition error is occurred in areas of the high speaking rate such as more than 6 morae/sec.

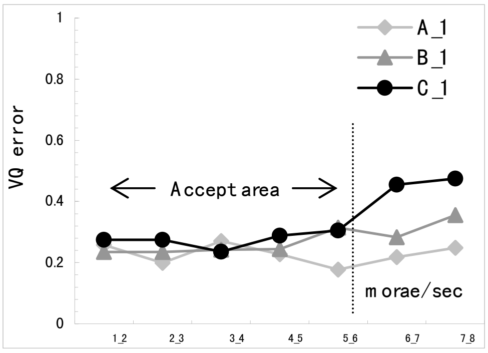

Fig.2 VQ errors before changing speaking rate

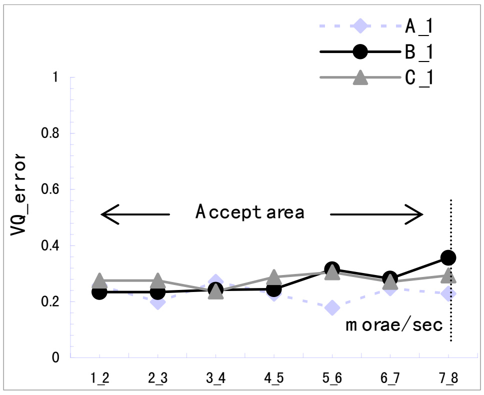

Fig.3 VQ errors after changing speaking rate

Especially the part of dialect distinction, to improve the distinction accuracy, this system is designed in the automatic response way. That is, the system works same the language identification when the VQ error is regular threshold δ_2 (it is less than 0.3 in Fig2), and chooses the high speaking rate sample codebooks automatically when the VQ errors are more than the δ_2. Fig.3 shows the other real test for the distinction of Mongolian dialects after changed samples .

3. Feature Selection

There might be many factors that affect identification performance acoustically as well as linguistically. It is crucial important that what kind of acoustic parameters are used for language and dialect identification at the high accuracy and the real time response.

In case of the speech recognition, the various parameters are usually chosen such as mel-cepstrum plus the delta mel-cepstrum. In this section, we investigate how the parameters for the language and dialect identification. Fig.4 and Fig.5 gave the VQ and language recognition errors used the various parameters and single one respectively on some languages. In case of Fig.4, with the increase of the speaking rate, the VQ errors increased sinificantly and the recognition error is easy to occur in areas of the low speaking rate, and this will be improved by using the mel-cepstrum such as figure.5. It is considered that the result in Fig.4 is caused by spectral transition speedily when the delta mel-cepstrum parameter was used[5].

Our system is designed to use the mel-cepstrum as performance parameter. It is favorable both to improve the recognition accuracy and the real time performance.

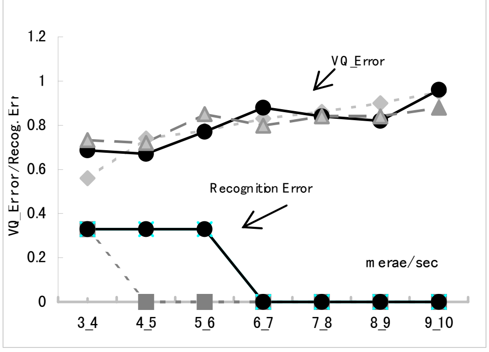

Fig.4 using mel plus delta mel-cepstrum

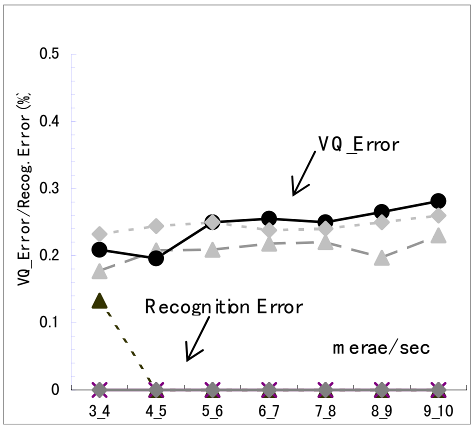

Fig.5 using mel-cepstrum only

4. SYSTEM EVALUATION

4.1 Speech materials

The system reported in this paper performed on the seven languages and four dialects of Mongolian. Table 1 shows the results. The speech materials of Korean and Thai are selected from Multilingual Speech Corpus produced by Tsukuba University[6], others are produced by Shirai laboratory of Waseda University in Japan. Speech sample data of 30–seconds spontaneous utterances for each of the languages were used, and test data is used in 1-8 morae/second utterances for each speaker.

4.2 Experimental conditions

Speech signals are digitized with 16kHz sampling and 16bit quantization. The digitized waveform is analyzed with 21.3 ms Hamming window, shifted by 5 ms interval. The centriod vectors are made by 15 dimensional LPC(*Linear Predictive Coding*) mel-cepstral coefficients with 256 points, and the test vector is used 15 dimensional LPC mel-cepstrum.

ISBN 1-876346-48-5 © 2003 UAB

Table1. Speech materials and test condition

Language	sample speaker 30-sec.	test speaker 1-8morae/sec.
Chinese	15	5×5
Japanese	15	5×5
Korean	10	5×5
Thai	10	5×5
Uigur	10	5×5
Khazak	10	5×5
Mongolian	30	10×5
Dialect	each dialect 10	5×5

4.3. Experimental results

Table.2 indicated the performance of demonstration test on our system. All of tests are performed by open-test (text independent). It can be confirmed in the table 2 that the misidentification rate on language is less than 3%, and the distinction error rate on dialect is approximately 6%.

Table.2 The system performance

Language	all tests	error	error rate(%)
	200	5	2.5
Dialect	100	6	6

5. CONCLUSION

In this paper, we introduced an application system for multilingual identification based on the VQ distortion, prosodic cue of speech and speaking rate . We described the process of the implementation of the system and its performances, and investigated some methods to improve the recognition accuracy and the real time response. The system is composed of a small number of parameters relatively, and its efficiency and robustness were confirmed by real experiments. We plan to extend languages in the usage and the number of speakers in our system quality in future.

ACKNOWLEDMENT

We thank Shuich Itahashi *et al,* for their support in the Multilingual Speech Corpus, and also appreciate deeply the cooperation of all speakers who help us record the speech data.

REFERENCES

[1] Idomuso Dawa, S.Okawa, K.Shirai “Design of Mongolian Speech Database considering Dialectal Characteristics”, J.Acoust. Soc.Japan,(E),May 1999.

[2] I.Dawa, K.Shirai, ”Development a real time automatic language identification”, ASJ,2002,9, pp161-162.

[3]Christina G.Foreman, ”DIALECT IDENTIFICATION FROM PROSODIC CUES”, ICPhS 99 Vol,2 pp 1237-1240.

[4] I.dawa , S.Okawa, K.Shirai “Analyzing and Classifying Mongolian Major Dialects by Acoustic and Prosodic Features”, Jornal of MINORITY LANGUAGES OF CHINA, 2001,1pp 26-31.

[5]Sadaki, Furui, ”On the role of spectral transition for speech perception”, J, Acoust, Soc, Am.80(4)Oectober, 1986, pp 1016-1025.

[6] S.Itahashi, *et al*, “Multilingual Speech Corpus”, Tsukuba University, 2002, CD-ROM.