A Multilingual Automatic Speech Recognition (ASR) Engine Embedded on Personal Digital Assistant (PDA)

Hong-wen Sie [1], Dau-Cheng Lyu [1], Zhong-Ing Liou [1], Ren-Yuan Lyu [1, 2], Yuang-Chin Chiang[3]
[1] Dept. of Electrical Engineering, Chang Gung University, Taoyuan, Taiwan
[2] Dept. of Computer Science and Information Engineering, Chang Gung University, Taoyuan, Taiwan
[3]Institute of Statistics, National Tsing Hua University, Hsin-chu, Taiwan
E-mail: rylyu@mail.cgu.edu.tw, Tel: 886-3-2118800ext5967

***Abstract*—**In the paper, we describe a multilingual ASR engine embedded on PDA. Our ASR can support multiple languages including Mandarin, Taiwanese and English simultaneously based on a unified three-layer framework, and a one-stage searching strategy. In the framework, there is a unified acoustic models for all the considered languages, a multiple pronunciation lexicon and a searching network, whose nodes represent the Chinese characters and English syllables with their multiple pronunciations. Under the architecture the system can not only reduce its memory and computational complexity but also deal with the issues about a character with multiple pronunciations. In general, the computer resource of PDA is quite limited when compared to PC. In this paper, much work has been done to alleviate the limitation of PDA. The experimental results show the system has good performance where the recognition rate achieves about 90% in the voiced command task with limited vocabulary.

1. INTRODUCTION

The main goal of this paper is to describe the task to improve the user-friendliness for the application in Personal Digital Assistant (PDA) based on a multilingual speech recognition engine. The PDA has many advantages, including small physical size and high mobility. However, the memory space and the computing power are not comparative to the personal computer (PC). Portability and low implementation resources are two key issues to determine which automatic speech recognition (ASR) solutions could be used in real-world embedded systems. A variety of ASR system architectures have been proposed for server based implementations [1-2]. In recent years, some applications of ASR in wireless mobile devices are also discussed [2-3].

Gradually, mobile phone devices and PDAs become indispensable for people in daily lives. It is probable that we can control all our electric appliances by using handled devices in the future. However, embedded systems have some shortcomings, such as small size of screen, the limited input modalities…, etc. Speech input will provide great convenient interface. But there are still some robustness issues to be solved. The reason is because the interference and channel noise of handled devices are more severe than that of PC. During the last decade, many researches have shown good result for speech recognition in noise environment. However it is still the main issue for speech recognition in the application to embedded systems.

In Taiwan, most people speak Mandarin, Taiwanese or even Japanese, English in their daily lives. It is very important for a speech recognition system to have the ability to deal with multiple languages. In Taiwan, Mandarin and Taiwanese usually co-exist in the daily conversion nowadays. They are even mixed in a single sentence. Although these two languages are quite different in spoken form, they share the same writing system. In our proposed multiingual (Mandarin, Taiwanese, English) system, more efforts were put to construct a multiple pronunciation lexicon proposed in the previous work [4]. In this paper, we consider three languages, i.e., Mandarin, Taiwanese and English, simultaneously within the same framework [5]. We have shown that this framework is also robust for application in the combinations of all the other Chinese regionalects and some Western languages.

In the market, there have been already many embedded products. Most of them used speaker-dependent technology. For such speaker-dependent systems, the users have to “train” the system by speaking a lot of utterances before using the systems. As the size of vocabulary becomes larger, it is even more difficult for the user to re-train the systems. To improve usability of these products, a multilingual speaker-independent speech recognition engine will be very useful.

This paper is organized as follows. In section Ⅱ, we describe the unified framework about the multilingual speech recognition engine. In sectionⅢ, a multilingual ASR system architecture on PDA is described. The experimental database information and results are reported in section Ⅳ, the paper ends with a summary of the results and future works.

2. THE UNIFIED FRAMEWORK OF MULTILINGUAL ASR

2.1. A Unified Framework for Multilingual Speech Recognition

Unlike the conventional approach which divides the recognition task as syllable decoding and character decoding, the new proposed approach adopt a one-state searching strategy as show in figure1, which decodes the acoustic feature sequence X directly to the desired character sequence C*, no matter what regionalects are spoken. The decoding equation can thus be shown as follows:

$$C^* = \arg\max_{C} P(X, C)$$

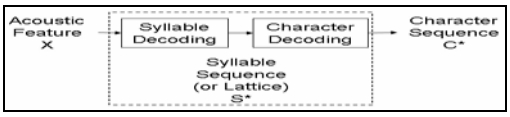

Fig 1 the decoding problem of Chinese speech recognition

In a unified framework, shown as figure2, there are at least 2 critical differences from the conventional one. One is in the lexicon layer, where the new framework adopts the character-to-pronunciation mapping which can easily incorporate the multiple pronunciations caused by multiple languages. For this purpose, we have used Formosa Lexicon [6]. Another one is in the grammar layer, where the character is adopted as the nodes of the searching net. This makes it be language independent! By the way, tone is a common feature for all Chinese languages. In the unified framework described above, it is easy to incorporate the tone feature into the system. There is no need to change the grammar layer. In another aspect, multilingual speech recognition is one of the most popular research topics recently in speech signal processing. It is essential to collect a large-scale multilingual speech database for research, especially for designing a speaker independent and multilingual speech recognition system.

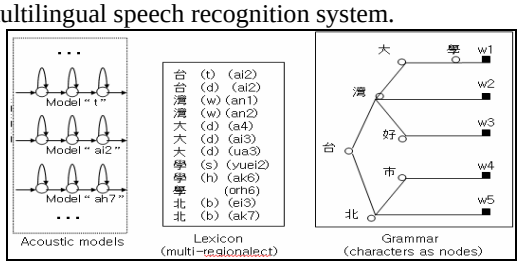

Fig 2 A Unified 3-layer frameworks for multi-regionalect Chinese speech recognition

2.2. Acoustic Modeling & Multiple Pronunciation Lexicon

The performance of any ASR system is highly dependent on the quality of the acoustic models. In particular, we need to support several languages in embedded systems at the same time. Due to memory size and capacity of embedded system are not bigger than that of PC, we adopted the Initials and tonal context-independent Finals as the HMM models. The same phoneme (including tone) symbols in the transcript in trilingual share the same training data. There are two approaches in dealing with pronunciation variations, i.e., knowledge-based approach and data-driven approach. The former consists of generating variants by using phonological rules, and the later consists of performing phone recognition to obtain information on the pronunciation variation in the data. We adopt the rule-based approach for Taiwanese tone sandhi and data-driven approach based on confusion matrix for finding phoneme mapping between the real pronunciations and canonical pronunciations. The matrix is constructed by a dynamic-programming technique to align the recognition results with the canonical transcriptions. We chose the most variational pronunciations by the confidence scores which is the occurrence possibility. The sum of all the occurred possibility of each character is then normalized to unity for fair competition in the Viterbi search.

2.3. Searching Strategies

A time-synchronous Viterbi beam search based on the token-passing algorithm was used as the basic searching algorithm in the new framework. Not like the entire Viterbi trellis search to find the optimal path, we use several pruning techniques to accelerate the searching speed by using certain heuristics. Common heuristic in beam search can be divided into the following categories:.

1. Using tree-based:

The previous papers [7] have pointed out the benefits of tree-based lexicon, especially in the large vocabulary speech recognition. It can merge the same token in the network, reduce the search space greatly, and prevent the memory squander while decoding.

2. Ranking the hypothesis:

In the time-synchronous searching, at each frame, only N (N=1000, for instance) active candidates are reserved for branching out in the next frame by ranking the acoustic scores.

3. Using probability score difference as threshold:

Any word hypothesis whose acoustic score is less than a threshold T (T=250 in log probability, for instance) from the maximal probability of this frame are discard. In other words, it is like hypothesis ranking, but it uses the score difference to keep the upper bound's candidates could branching out in the next frame.

4. Using level constraint width:

It's a combinational strategy of the second and the third heuristics mentioned above. During the search, the word hypotheses, which meet the above two requirements simultaneously are kept alive. Therefore, the number of hypotheses is a function of the number of symbol levels recognized, just like the dynamic threshold for searching in a net. We believe that due to the tree-based lexicon, there is a strong constraint in the searching token. It's reasonable to keep more hypotheses alive in the initial because every token has opportunity to compete against others. Nevertheless, the more close to the end of the search, the more token restriction there is. For this reason, we should prune more hypotheses in the tail of the search.

3. MULTILINGUAL ASR SYSTEM ARCHITECTURE ON PDA

3.1. System Architecture

Based on the framework mentioned above, the multilingual ASR engine consists of two parts: multiple pronunciation lexicon and on-line pronunciation acoustic models. In our system, the acoustic models are shared for each language. With our multiple pronunciation lexicons, the system can automatically recognize multilingual vocabulary items without the language identification. Figure3 illustrates the architecture for multilingual ASR system. With our multilingual, speaker independent and large vocabulary speech engine, speech signals can be converted into Mandarin, Taiwanese or English. Therefore, the results can lead to applications of the lower level, such as: mp3 broadcast machine, voice name dialing and a turntable of TV program in our system.

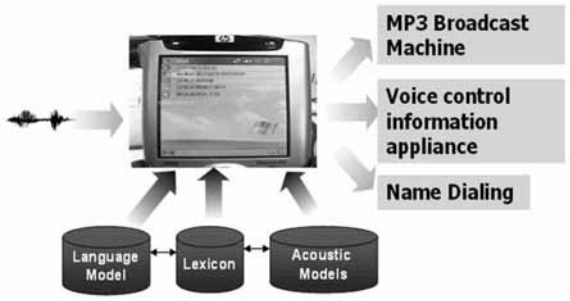

Fig 3 Application procedure of system based on multilingual speech recognition

3.2. The Requirement of System Resource

Resources we need in our system shown as table 1. In the second column, it represents the usage of capacity that the program is not executed by user, e.g. 84.0KB is needed by system program in storage card. In the other hand, the usage of capacity system needs is shown in the third column when the user executes our system on PDA.

STORAGE POSITION / Domain	Storage Card	Memory
Program	84.0KB	4.3MB
Models and Lexicon	2.95MB	6.22MB
Total	3MB	10MB

Table 1 The table states resources which our system needs.

We just need 3MB memory to storage our acoustic models and multiple pronunciation lexicons, and 10MB memory size is located when executes it on PDA. Therefore, our system can be loaded in almost embedded systems.

3.3. The Issue of System Interference

In our system, we find out some problems that the hardware will produce interference noise when we press the record botton during recording voice. This problem reduces recognition performance very much and we cut off head and tail of about 0.5 second audio signals to solve this problem since the noise is in fixed time. Although the channel noise is much on PDA more than PC, this problem, which was produced by microphone, can be overcame with our robust acoustic model. This interference and channel noise shows as figure4.

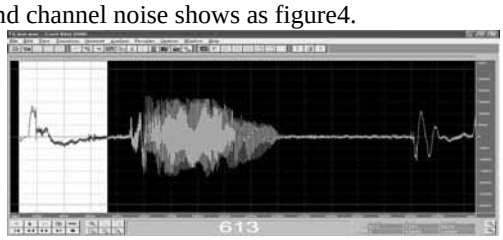

Fig 4 Hardware interference and channel noise of PDA, 0~0.5 sec is hardware interference, 0.5~1.2 sec is recorded speech, the last 0.3 sec also are channel noise and hardware interference.

4. EXPERIMENTS AND RESULTS

4.1. Database Information

In our multilingual speech recognition system, our database used for the experiments are given in Table2. ForSDat has been collected over both microphone and telephone channels by Chang-Gung University, namely, ForSDat-TW01, ForSDat-MD01, ForSDat-TW02 and ForSDat-TW03, respectively. In addition, the multilingual lexicons (Formosa Lexicons) are also important parts for building the corpus. The tag "TW01" means that a portion of the database was collected in 2001 in Taiwanese. On the other hand, the tag "M0" means that the recording channel used was a microphone and gender was female, and so on. Every speaker has a unique serial number and speech data, which contain a transcription of waveforms, made in the early stage and are stored in a unique folder named according to the serial number. All the statistics of the database are listed in Table 2.

	Name	Channel	Gender	Quantity	Train (hr)	Test (hr)
ForSDAT	TW01-M0	MIC	Female	50	5.92	0.29
	TW01-M1		Male	50	5.44	
	MD01-M0		Female	50	5.65	0.27
	MD01-M1		Male	50	5.42	
	TW02-M0		Female	233	10.1	0.7
	TW02-M1		Male	277	11.66	
	TW03-M0		Female	409	74.61	
	TW03-M1		Male	264	45.56	
	TW02-T0	TEL	Female	580	29.21	0.95
	TW02-T1		Male	412	19.37	

Table 2 The statistics of utterances, speakers and data length for speech collected over microphone and telephone channels in Taiwanese and Mandarin (MIC: microphone; TEL: telephone).

4.2. The Performance of PC Based System

The baseline recognition system was trained by using a bi-lingual Mandarin/Taiwanese speech corpus. The baseline system is to show performance of our proposed unified framework. In the Tang Poem Task, the testing data set contains 2500(1500 Taiwanese and 1000 Mandarin) utterances recorded by 4 males, and 3223 sentences with 16865 characters. In table 3, the results shows that under the same probability score difference (T=200), the character accuracy rate (92.5%) is 1%

lower than the case in (T=300). However, the speed is much faster than the case in (T=300), therefore the level constraint width is a better beam search strategy to minimize the computational cost with a minimal decrease in recognition accuracy.

	CAR (%)	Speed (xRealtime)
(1)+(2)+(3)+(T=200)	88.4	1.37
(1)+(2)+(3)+(T=300)	93.5	4.26
(1)+(2)+(3)+(4)+(T=200)	92.5	1.75

Table 3 The character accuracy rate (CAR) of the task of the Tang-poem, under different combinations of searching strategies: (1). Using tree-based lexicon, (2), ranking the hypothesis, (3). Using probability score difference, (4). Using level constrain width.

4.3. The Performance of PDA System

On PDA, we use TW01 and MD01 as our training corpuses. Our system was established on the language-independent sets, with phoneme sharing between languages. The audio signal is 16 kHz, and 39-dimensional MFCC features were computed. In order to test performance of our system on PDA, we make a simple experiment to understand effect on PDA. The test set, contains three kinds applications, contains 442 sentences (about 904.5 seconds) recorded from 1 speaker. In every kind, we make a statistical table about sentences of different language. It is shown as Table 4. The second column is sentences of Mandarin. For example, these names of song are 10 sentences belong to Mandarin.

Language / Domain	Sentence of Mandarin	Sentences of Taiwanese	Sentence of English
TV	30	25	24
Name	68	20	15
Song	102	75	51

Table 4 Statistics about sentences of difference language in four domains

The test results shown in Table 5 are promising. The result was close to the monolingual system. The first column is application domain, e.g. the “TV” represents TV name. The second column is numbers of entire for speech recognition engine. In the forth column, SER is sentence error rate and the forth row is the lowest SER. This is because of the same first name in our testing names. It can also be seen that the length of speech doesn’t affect recognition time. The entirety velocity majors with the amount of vocabulary. Although the amount of vocabulary is not many, but this situation in lives of people is enough. The SER of the third is 94.93% which is high enough to accept it. By observing Table5, we find out the recognition time is longer in our system. This fact reveals fixed-point operation is not applied and PDA do not support float-point accelerator yet. In another aspect, our ForSDat database is recorded on notebook at that time. Acoustic models and testing set isn’t the same condition of environment that is possible a factor which affects our recognition rate.

Domain	Vocabulary	Length (s)	SER(%)	Time (s)
TV	79	2.431	94.93	23.57
Name	103	1.569	84.87	35.02
Song	228	2.087	88.60	90.92

Table 5 The statistics of our system performance are vocabulary, length, time and SER (Sentence Error Rate)

5. CONCLUSIONS AND FUTURE WORKS

In this paper, we have investigated that the unified framework is robust when combining 2 Chinese regionalects, including Mandarin and Taiwanese. It seems still workable when some English words were added into the vocabulary list. Under the unified framework, we achieved promising results on PDA. We not only saved the computing resource required to perform the speech recognition engine but also provided multilingual ability in this framework. From these experiments, we have shown that the recognition rate is still acceptable when the engine was ported from PC to PDA. At present, the recognition rate is almost 80%~90%. Although the recognition speed is slow, this problem will be solved. In future, multilingual support will be a trend on handled devices. It is easy to use this framework to integrate more languages, such as Hakka, Cantonese, Shanghais and so on. This will truly increase the user-friendliness for the Chinese society.

REFERENCES

[1] S. H. Maes et al. “Conversational networking conversational protocols for transport, coding, and control,” Proc. Int. Conf. on Spoken Language Processing, October 2000.

[2] R. C. Rose et al. “On the implementation of ASR algorithms for hand-held wireless mobile devices,” ICASSP ’01.

[3] Olli Viikki, “Asr in portable in wirelss devices,” Automatic Speech Recognition and Understanding, 2001. ASRU ’01.

[4] Dau-Cheng Lyu et al. "Large Vocabulary Taiwanese (Min-nan) Speech Recognition Using Tone Features and Statistical Pronunciation Modeling," In Proc. EuroSpeech 03, Geneva, September 2003

[5] Ren-Yuan Lyu et al. “A Unified Framework for Large Vocabulary Speech Recognition of Mutually Unintelligible Chinese Regionalects,” ICSLP,2004.

[6] Min-siong Liang et. al. “Construct a Multi-Lingual Speech Corpus in Taiwan with Extracting Phonetically Balanced Articles,” INTERSPEECH 2004 - ICSLP, Jeju island, Korea.

[7] Hermann Ney, Ortmanns. S, “Progress in Dynamic Programming Search for LVCSR,” Proceedings of the IEEE, Vol. 88, pp. 1224-1240, August 2000.

[8] Xia Wang et al. “An Embedded Multilingual speech recognition system for Mandarin, Cantonese, and English” In Proc. Natural Language Processing and Knowledge Engineering, October 2003.

[9] Jyh-Shing Roger Jang, Shiuan-Sung Lin, "Optimazation of Viterbi Beam Search Speecg Recognition," In Proc. Internal Symposium on Chinese Spoken Language Processing, Taipei, August 2002

[10] Mingkuan Liu et al. "Mandrain Accent Adaptation Based on Contest-Independent/Context-Depent Pronunciation Modeling," In Proc.ICASSP 00, 2000

[11] Mirjam Wester, "Pronunciation Modeling for ASR-knowledge-based and Data-drived Methods," Journal of Computer Speech and Language 17(2003), pp. 69-85, 2003