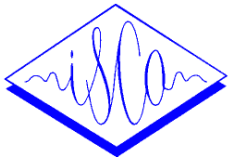

International Symposium on Chinese Spoken Language Processing (ISCSLP 2002)
Taipei, Taiwan
August 23-24, 2002

EFFICIENT PHONE-BASED RECOGNITION ENGINES FOR CHINESE AND ENGLISH ISOLATED COMMAND APPLICATIONS

Xavier MENÉNDEZ-PIDAL, Lei DUAN, Jingwen LU, Beatriz DUKES, Michael EMONTS, Gustavo HERNÁNDEZ ÁBREGO, Lex OLORENSHAW
Spoken Language Technology Group, SONY NSCA,
San José, CALIFORNIA
{Xavier, Lei, Jingwen, Beatriz, Mike, Gustavo, Lexo} @slt.sel.sony.com

ABSTRACT

In this paper we present a flexible and efficient approach to perform an accurate speech recognition interface for isolated command applications in three different languages: Mandarin, Cantonese and English. The paper analyzes and discusses the different trade-offs necessary to obtain an accurate, real-time system with low memory requirements. Areas addressed are design of the training database, and Hidden Markov Model (HMM) units used by the recognizer (monophones versus triphones).

1. INTRODUCTION

A speech recognition interface to control small computer devices like command car navigation, telephone, or robot systems is discussed in this paper. In such devices with limited CPU and memory resources, the implementation of a flexible real-time and accurate speech recognition interface is an open engineering design issue. The 2 restrictions in our recognizer, memory and maximum computational cost, are introduced next. For the system analyzed, we had a maximum memory size of 0.5 Megabytes. Speech recognition is a cumbersome process even for small command applications, but can be accelerated by pruning techniques such as beam search strategies in a standard Viterbi search engine. During the Viterbi decoding no more than 300 Gaussians per frame can be estimated to meet the CPU limitations of our usual hardware. In the paper we analyze and compare 2 recognizer designs based on monophone and triphone models to obtain a real-time system with limited memory requirements. Also, we discuss different alternatives to improve system accuracy and portability using an appropriate training and testing database design.

2. SYSTEM DESCRIPTION

The speech recognition systems investigated were based on basic phone-like units (PLUs) to obtain a task independent system. These systems can be used easily in different command control applications by changing the dictionary description in each new task. The English system is based on 43 phone units. The Mandarin is based on 37 PLUs covering both northern and southern Mandarin dialect variations. The Cantonese system uses 38 basic phones. Neither the Cantonese nor the Mandarin system models tones in order to simplify the system implementation.

The recognizer front-end was based on 20 Mel Frequency Cepstral Coefficient (MFCC) features to provide a good trade-off between accuracy and computational cost. The front-end includes the first 6 cepstral features (C1-C6), their velocity and acceleration features, and the velocity and the acceleration of the C0 feature. No channel normalization procedure was applied.

The monophone and triphone recognizers analyzed in this research were based on standard 3-state continuous HMM systems that tend to provide the maximum accuracy. In the monophone system the ½ Megabyte memory restriction can be met using HMMs with less than 25 Gaussians per state for each language. In the Tree-based clustered triphone

system the best system configuration was obtained using a 1-Gaussian per state HMM with a control amount of different states. For the triphone recognizer, the memory requirement was met using 1 Gaussian per state with: 1500 different states for the Mandarin and Cantonese systems and 1400 states in the English recognizer. Using larger HMM configurations, a much higher accuracy can be obtained, but they usually need more memory and are computationally more expensive. For the 1-Gaussian per state triphone system the real-time CPU restrictions are met in our hardware by using a beam search of 300 active states.

3. TASKS AND DATABASES DESCRIPTION

In some preliminary experiments in English we observed that training with only general data does not produce an efficient system on a test set when the test triphones are not properly covered in the training set. To obtain an accurate but general system, task specific data was included to cover the triphone set of our target application. Whenever possible, a general-purpose training set was added in the training corpus to improve system portability and generality.

In the English system 20 hours of general purpose isolated and continuous speech uttered by 200 speakers was combined with 5 hours of robot application task-specific data uttered by 14 to 28 speakers, depending on the task. In the Mandarin system we combined the ASCCD data set [1,2] comprised of about 9 hours of read speech produced by 10 speakers with 4 hours of task specific speech produced by 35 speakers (17 Northern + 18 Southern Mandarin Speaker) recorded in our facilities. In the Cantonese system we experimented with combining the generic CUWORD database [3,4] with our training data set, comprised of 3 hours of speech uttered by 20 speakers. The Mandarin and Cantonese training and testing sets are composed respectively of 350 (ER350M) and 250 (ER250C) commands related to entertainment robot (ER) applications. In order to cover all the Chinese syllables, 120 general Mandarin phrases and 210 Cantonese general greetings, were recorded. The English testing set tasks are 3 entertainment robot applications related to the Sony's pet robot (ERC100, ERA250, ERS125) and a car navigation command task (CAR125). The first two tasks (ERC100, ERA250) have 100 and 250 commands respectively. In the ERC100, ERA250 and Chinese tasks, many commands match the same target action. We optionally reduce the original vocabulary size to improve system accuracy [5,6], while at the same time cover all the target actions required by the system. The ERS125 (more related to the final AIBO Pal application than other ER tasks), and the CAR125 task (a car navigation task), both have a vocabulary size of 125 commands. In the Mandarin test set we recognized two tasks: one with 350 AIBO commands (ER350M), and a second task including all 470 commands (All470M). Similarly, In the Cantonese system two sets were tested: one with 250 commands (ER250C), and the whole set of 460 commands (All460C).

In each training and testing set, there is a balance between male and female speakers with no overlap between training and testing speakers. The number of test speakers is not uniform across the tasks and varies from 8 to 16 speakers. In English: ERC100 has 8 speakers; ERA250: 8 speakers; CAR125: 15 speakers; ERS125: 14 speakers. In Mandarin, the testing set has 16 speakers (8 from southern China and 8 northern). Finally, the Cantonese testing set is composed of 12 speakers from Hong Kong. In the Mandarin test sets some speakers with very strong dialects were discarded. Mostly, the Chinese speaking style was very close to formal read speech. Speaking style in the English robot task sets was more informal. Some speakers exhibited a more emotional speaking style, reflecting the expected emotional interaction of a robot owner with his SONY pet robot.

4. DATABASE DESIGN AND SYSTEM ACCURACY

In this section, we analyze the influence of the training corpus to obtain an accurate and general triphone HMM system. Table 1 shows the different system accuracies for the English triphone system trained on different training corpora. The *Gen* system

includes only a general training database. The *ETask+Gen* system was trained with robot task-specific data and general data. The *Etask* system was trained only with task-specific American English data. In this experiment we see the benefit of mixing a large database of generic data with a small database of task-specific data to cover all the triphones of a target application to obtain an accurate and portable system in unknown scenarios.

Train / *Test*	*Gen*	*Etask+Gen*	*ETask*
ERC100	92.3	95.1	95.7
ERA250	88.7	95.2	94.0
ERS125	94.2	96.0	97.2
Ave	**91.7**	**95.4**	**95.6**

Table 1. Word accuracy in the English system using different training sets

In Table 2 we show results obtained when mixing a general-purpose training database with a task-oriented training database for the Mandarin (MT+ASCCD) and Cantonese (CT+CUWORD) triphone systems. It also shows results obtained for the Mandarin (MT) and Cantonese (CT) systems trained on robot task data alone, recorded at Sony.

Train / *Test*	*MT+ ASCCD*	*MT*	*CT+ CUWORD*	*CT*
ER350M	98.3	98.4		
All470M	97.7	97.5		
ER250C			92.2	96.2
All460C			92.6	96.9
Ave	**98**	**98**	**92.4**	**96.6**

Table 1. Word accuracy in the Mandarin and Cantonese systems using different training sets

The Mandarin and English systems behave similarly when the task-specific and general-purpose databases are used together to train the systems. On the other hand, the Cantonese CUWORD database greatly decreased our system accuracy. The decrease in accuracy introduced by the CUWORD database seems to be due to channel or microphone mismatch between the CUWORD recording environment and ours. While the ASCCD database sounds acoustically similar to our databases, the low frequencies seem to be boosted in the CUWORD database. This mismatch can be compensated using a channel or database normalization scheme in the front-end, but has not been used thus far.

5. MONOPHONE AND TRIPHONE SYSTEM COMPARISON

In Table 3 we show the accuracy of the Mandarin and English monophone recognizers for different system configurations. In the monophone system, the memory restriction can be met using HMMs under 25 Gaussians per state. Very good system performance for the monophone system is usually obtained by using 8 to 10 Gaussians per state. In Table 3, results with a good monophone system using wide beams (600) and 16 Gaussians are illustrated. On the other hand, CPU limitations tend to degrade system accuracy of the monophone HMMs, as shown by the right-most column in Table 3. This column displays results for monophone HMMs with 2 Gaussians per state using a beam search of 150 active states.

	16Gaussians-Beam600 (340Kb)	2Gaussians-Beam150 (45Kb)
All470M	97.1	88.1
ERA250	95.0	76.4
ERC100	93.9	77.8
ERS125	95.3	76.5
CAR125	98.6	88.3
Ave	**96**	**81.4**

Table 3. Word accuracy in the Mandarin and English monophone HMM systems

Though the 2-Gaussian monophone systems are able to run at real-time on our platform, their accuracy

demonstrates the limitations of the monophone architecture. Since a system with such accuracy is not suitable for a product, triphones are considered.

In Table 4 we summarize the final average results obtained with 3 triphone systems in the three languages over all the tasks. These results were obtained with 1/2 Megabyte HMMs with 1 Gaussian per state using a narrow beam search of 300 states.

	English	Mandarin	Cantonese
Ave	**95.5**	**98**	**98.2**

Table 4. Average Word accuracy in the Mandarin, Cantonese and English triphone HMM systems

The English tasks are more difficult applications. In all the tasks there are many commands phonetically close with completely different meanings: bed/beg, no/go, go right/good night, bark/back, etc. The English system can be improved to 97% word accuracy by applying command selection to eliminate commands that are easily misrecognized. The Cantonese system was improved from 96% to 98% word accuracy by merging all the commands with same meaning. Most of the recognition errors in the Cantonese system were due to commands with a different final particle but with the same meaning. All those commands were grouped together to improve accuracy. The Mandarin system was improved by simply using a dictionary where entries with free allophonic and dialect variations were all merged together.

6. CONCLUSIONS

In this paper we analyzed technical design issues to obtain an efficient speech interface for command control applications with low CPU cost and limited memory requirements. Traditional phone-based speech recognizers can be used for Western or Chinese languages to implement a successful command recognition application. Triphone recognition engines are more CPU-efficient than monophone systems. To obtain an accurate speech interface based on triphones, the target triphones of applications need to be well covered in the training corpus. Good triphone coverage can be obtained by recording the target vocabulary with 15 to 30 speakers. The use of a general-purpose database may improve system portability to an unknown scenario, but may not help the target application: 1) if channel or microphone characteristics diverge severely or, 2) if the triphones of the target application are already well covered in the training corpus. For recognizers with limited resources, vocabulary design is still a major concern. To obtain an accurate system, commands that are phonetically close should be avoided.

7. ACKNOWLEDGEMENTS

Special thanks to our linguistic technicians for recording and transcription: Michelle Chen, Eve Culver, Cameron Ladd, Carey Little, Salman Mattu, Sarah Pierce, Ariana Stamper-Gimbar, Marge Sung, Greg Uchishiba, Hideki Yamashita

8. REFERENCES

[1] LI Aijun, Zheng Fang, William Byrne, Pascale Fung, Terri Kamm, LIU Yi, CHEN Xiaoxia, "CASS A Phonetically Transcribed Corpus of Mandarin Spontaneous Speech," *Proceedings of International Conference on Spoken Language Processing*, Beijing, China, 2000, pp. 485-489.

[2] LI Aijun, CHEN Xiaoxia, Sun Guahua, Hua Wu, Yin Zhing, ZHENG Fang, SONG Zhanjiang, "The Phonetic Labeling On Read And Spontaneous Discourse Corpora," *Proceedings of International Conference on Spoken Language Processing*, Beijing, China, 2000, pp. 724-727.

[3] W.K LO, Tan LEE and P.C. CHING, "Development of Cantonese Spoken Language Corpora For Speech Applications" *Proceedings of the 1998 International Symposium on Chinese Spoken Language Processing*, Singapore, 1998, pp. 102-107.

[4] W.K LO, K.F. CHOW, Tan LEE and P.C. CHING, "Cantonese Databases developed at CUHK for Speech Processing" *Proceedings of the Conference on Phonetics of the Languages in China*, Hong Kong, China, 1998, pp. 77-80.

[5] X. Menéndez-Pidal, G. Hernández Ábrego, L. Olorenshaw, "Optimal Command Selection for the AIBO Speech Interface", Technical Digest 11th Sony Research Forum, 2001, Tokyo, Japan, pp 161.

[6] G. Hernández Ábrego, X. Menéndez-Pidal, L. Olorenshaw, "Robust and Efficient Confidence measure for Isolated command application", *in Proceedings of Automatic Speech Recognition and Understanding Workshop*, Madonna di Campiglio, Trento, Italy, December 2001.