行政院國家科學委員會專題研究計畫　期中進度報告

中文口語處理技術之前瞻性研究課題(1/3)

計畫類別： 個別型計畫
計畫編號： NSC91-2219-E-002-040-
執行期間： 91 年 08 月 01 日至 92 年 07 月 31 日
執行單位： 國立臺灣大學電信工程學研究所

計畫主持人： 李琳山

報告類型： 精簡報告
報告附件： 出席國際會議研究心得報告及發表論文
處理方式： 本計畫可公開查詢

中　華　民　國 92 年 6 月 13 日

行政院國家科學委員會專題研究計畫期中報告

中文口語處理技術之前瞻性研究課題(1/3)

計畫編號：NSC 91-2219-E-002-040

執行期限：91 年 8 月 1 日至 92 年 7 月 31 日

主持人：李琳山　　國立台灣大學電信工程學研究所

E-mail: lslee@gate.sinica.edu.tw

ABSTRACT

With the rapid developments of wireless communications, it is highly desired for users to access the network information with spoken dialogue interface via hand-held devices at any time, from anywhere. One possible approach towards this goal is to perform speech feature extraction at the hand-held devices (the clients) and have all other recognition tasks and dialogue functions absorbed by the server. This report investigated distributed Chinese keyword spotting and verification under this scenario. A “phonetically distributed” Mandarin speech database including all possible Mandarin syllables and context relationships with frequencies roughly proportional to those occurring in daily Mandarin conversation is used to train a best set of vector quantization codebooks, such that the syllable recognition accuracy degradation due to quantization errors is minimized. Enhanced Chinese keyword spotting techniques were then developed using utterance verification approaches with weighting parameters optimized by MCE training. Experimental results indicated that the keyword verification approach achieved significant improvements in keyword spotting performance, and the overall results integrating vector quantization, keyword spotting and verification is quite satisfactory.

中文摘要:

隨著無線通訊的快速發展，我們將可以透過手機等攜帶方便的工具提供的語音介面，隨時隨地取得想要的資訊或知識。想要達到這個目標，一個可能的做法是，在使用者手上的機器這邊，做好抽取語音特徵參數的工作，剩下的工作，諸如語音辨識或是對話介面，都交給遠端的伺服器執行。本篇論文主要探討的就是，在這種環境之下的中文關鍵詞檢測與認證。訓練語料是一套包含所有中文音節以及其前後關係和頻率資訊的資料庫，使用這樣的語料庫訓練一套最佳的向量量化碼本，以極小化量化錯誤對語音辨識正確率的影響。接著提出強化的中文關鍵詞檢測的方法，概念是語音認證的技術，再加上最小分類錯誤訓練法訓練出來的參數。

由實驗結果可以看出關鍵詞認證對關鍵詞檢測的確有顯著的幫助，而且整體向量量化、關鍵詞檢測與認證的結果皆令人滿意。

1. INTRODUCTION

With the rapid development of wireless communications, strong demands are emerging for the users to access the network information via simple hand-held devices at any time, from anywhere. As the size of such hand-held devices shrinks while that of human fingers doesn't, spoken dialogues will no doubt become one of the few most attractive interfaces for user-network interaction. The concept of Distributed Speech Recognition (DSR) was developed under such a scenario, in which the complicated speech recognition processes can be distributed over the network with servers absorbing the majority of the computational requirements and the hand-held devices as the clients.

One possible approach for such DSR concept is to have the speech signal feature extraction performed at the hand-held devices (or the client), while leaving everything else at the servers. There are quite several possible advantages for this approach. First, the computational requirements at the client are low, thus practically feasible for many different designs of hand-held devices. Second, transmission of the compact feature parameters via wireless channels may need much less bandwidth, which is very precious for wireless networks, as compared to the transmission of the entire speech signals. Of course, there also exist various difficulties for this approach as well. One of them is the efficient representation of the feature parameters in form of binary digits, so that they can be transmitted in wireless packets. The Split Vector Quantization (SVQ) approach as proposed by the European Telecommunication Standards Institution (ETSI) is one possible approach, in which the MFCC parameters are vector quantized before transmission. But one challenging issue, among many others, with this approach is the extra recognition errors induced by the quantization errors. The work described in this report is based on the SVQ approach, but considering the phonetic distribution of Mandarin Chinese.

The focus of this report is two-fold. The first is the SVQ codebook design taking into account the phonetic distribution of Mandarin syllables. The purpose is to make sure the degradation in accuracy for Mandarin syllable recognition can be minimized after the vector quantization. This is because Mandarin Chinese is a syllable-based language and syllable recognition accuracy translates directly into any other metrics for user-network interaction, such as the word accuracy, keyword spotting rate, average precision rate for information retrieval, or speaker intention understanding rate. The second focus of the paper is the enhanced approaches for Chinese keyword spotting, because keyword spotting is the enabling fundamental element for many spoken dialogue systems. Here we used improved utterance verification approaches with parameters optimized by Minimum Classification Error (MCE) methods to verify the spotted keywords. In other words, the input keyword is accepted if

$$\phi(X \mid k) > th \qquad (1)$$

else we reject it, where X is the observed signal, k is the spotted key-word, $\phi(\cdot)$ is some confidence measure, and th is a threshold. This implies the primary goal is to keep the false alarm rate very low. This is because incorrectly received keywords may lead to misunderstanding in dialogues, which is difficult to recover, therefore such errors should be minimized in any case. On the other hand, if some keywords are missing, the system can always ask the user to repeat his question. This is the way to improve the dialogue reliability.

2. SPLIT-VECTOR QUANTIZATION (SVQ) FOR DISTRIBUTED SPEECH RECOGNITION (DSR) ENVIRONMENT

A Distributed Speech Recognition (DSR) system is one in which the network clients may call for some recognition services and the recognition task is primarily performed in the server, while the network may include wireless channels or the Internet. One of the most significant issues in DSR is to transmit sufficient speech information for recognition purposes in a band-limited channel and to minimize the recognition errors, given unavoidable quantization errors and channel transmission errors. Besides, because the client devices may have very limited computation functionalities and memory size, a proper division of the recognition task between the client and the server is also very important. In this report, our approach was based on the ETSI standard [1]; that is, at client side the MFCC (Mel-frequency Cepstrum Coefficient) coefficients are vector quantized and then transmitted to the server, and the first and second derivatives for MFCCs are then generated and the recognition task performed at the server. In this way, the limited computation requirements for feature extraction makes it feasible for many different designs of hand-held devices to serve as the clients, and all the complicated speech recognition and dialogue modeling jobs can be absorbed by the server.

We used 13 MFCCs per frame, consisting of C_1, C_2,…, C_{12} plus the log-energy E. Without vector quantization, the data rate required for transmission of the 13 real numbers is still high as high as 41.6 kbps(1 floating number = 32 bits, a total of 13 floating numbers per frame, 100 frames per second). This is why we need to compress the MFCC vectors by SVQ (Split Vector Quantization) as proposed by ETSI, although the detailed quantization scheme used in the studies here are slightly different. The procedure for SVQ is simple. Fist, a few coefficients are grouped together to form sub-vectors, and then a VQ codebook is trained for each sub-vector. Based on the ETSI experiences, we assume that neighboring MFCCs (eg. C_n and C_{n+1}) are usually more correlated, while the log-energy component is less correlated to all the other components. So we group every two subsequent MFCCs into a sub-vector, and leave the energy component alone to be a single sub-vector. This is illustrated in Figure 1. We then assigned different codebook sizes for each sub-vector considering the desired transmission bit rates; that is, the allowed quantization errors for different sub-vectors should be properly selected. Finally, we use k-means algorithm [2] to train the codebooks for the sub-vectors. The detailed experimental results will be further illustrated later on in section 4.

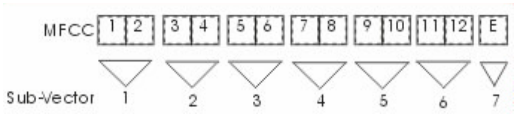

Figure 1: Building up sub-vectors.

3. KEYWORD SPOTTING AND VERIFICATION

In a dialogue system, the incorrectly recognized keywords may lead to serious misunderstanding. In such cases the user may need to use very long conversation to bring the dialogue back to the right track, if the conversation doesn't totally deviate from what the user wants. But if some keywords are missing, the system can always ask the

user to repeat the question. Such situation becomes more critical in wireless environment. This is why spotting and verification of keywords are important.

With the above considerations, the approach used in this report is key-word verification based on utterance verification approaches. The simplest utterance verification is the so-called dichotomizer, i.e., to use the likelihood ratio of the best key-word candidate to the 2nd key-word candidate to determine the confidence measure of the best key-word [3],

$$°(X \mid k^{(1)}) = ¢[\frac{p(X \mid k^{(1)})}{p(X \mid k^{(2)})}] > th \qquad (2)$$

where $¢[.]$ is the sigmoid function, $k^{(1)}$ is the best keyword candidate, $k^{(2)}$ is the 2nd key-word candidate, X is the observed signal, $°(X \mid k^{(1)})$ is the confidence measure for the best keyword $k^{(1)}$, and th is a threshold. Therefore the best keyword candidate $k^{(1)}$ is accepted when the confidence measure is high enough; else we reject it.

There have been a variety of different approaches to verify the utterances, just to name a few below.

- Using anti-models trained from the cohort set Q of the desired keyword [4],

$$°(X \mid k^{(1)}) = ¢[\frac{p(X \mid k^{(1)})}{p(X \mid \bar{k}^{(1)})}] \qquad (3)$$

  where $\bar{k}^{(1)}$ is the anti-model of $k^{(1)}$.

- Using the posterior probability directly as the confidence score [5],

$$°(X \mid k^{(1)}) = p(k^{(1)} \mid X) = \frac{p(X \mid k^{(1)})p(k^{(1)})}{p(X)} \qquad (4)$$

- Using filler models and OOV models with the N-best word candidates [6],

$$°(X \mid k^{(1)}) = \sum_{i=1}^{N} a_i ¢[\log(\frac{p(X \mid k^{(i)})}{p(X \mid \bar{k}^{(i)})})] \qquad (5)$$
$$+ a_¤ ¢[\log(\frac{p(X \mid ¤)}{p(X \mid \bar{¤})})] + a_{¨} ¢[\log(\frac{p(X \mid ¨)}{p(X \mid \bar{¨})})]$$

  where $k^{(i)}$ is the i^{th} best key-word candidate, $¤$ is the filler model, $¨$ is the OOV model, $\bar{k}^{(i)}$, $\bar{¤}$ and $\bar{¨}$ are the anti-models, $a_¤$, $a_{¨}$, a_i are weighting coefficients to be selected, and the summation in the first term is over all the N-best candidate words.

The key-word verification approach used in this report is similar to that in equation (5), but with a small modification as given below, which was found to provide good results for the task here in our experiments.

$$°(X \mid k^{(1)}) = \sum_{i=1}^{N} a_i ¢[\log(\frac{p(X \mid k^{(i)})}{p(X \mid \bar{k}^{(i)})})] + a_¤ ¢[\log(\frac{p(X \mid ¤)}{p(X \mid \bar{¤})})] \qquad (6)$$

where the likelihood scores for $\bar{k}^{(i)}$ and $\bar{¤}$ were approximated by the geometric mean of the relevant probabilities, as shown in the following formula,

$$p(X \mid \bar{\lambda}^{(i)}) = [\prod_{j=1, j\neq i}^{N} p(X \mid \lambda^{(j)})\, p(X \mid \sigma)]^{1/N} \qquad (7)$$

and

$$p(X \mid \bar{\sigma}) = [\prod_{j=1}^{N} p(X \mid \lambda^{(j)})]^{1/N}. \qquad (8)$$

The coefficients $\{a_1, a_2 \ldots, a_N, a_\sigma\}$ were optimized using the Minimum Classification Error (MCE) method, as proposed previously [6].

4. EXPERIMENTAL RESULTS

The first task is to develop a set of VQ codebooks good enough in minimizing the syllable recognition accuracy degradation due to quantization errors for Mandarin Chinese. This is because Mandarin Chinese is a syllable-based language, and syllable recognition accuracy translates directly into any other metrics such as word accuracy, keyword spotting rate, average precision for information retrieval, and speaker intention understanding rate. A carefully designed "phonetically-distributed" Mandarin speech corpus for daily conversation applications was therefore used in training the VQ codebooks. This corpus not only includes all the Mandarin syllables and context relationships, but all the Mandarin syllables and context relationships appear in this corpus with frequencies roughly proportional to their distributions in daily conversation in Mandarin Chinese. The VQ codebook optimized with this corpus therefore serves the purpose here in this study.

VQ Codebooks	Bit Rates(kbps)	Bit Allocation to the sub-vectors						
(A) 8777766	4.7	8	7	7	7	7	6	6
(B) 7766666	4.4	7	7	6	6	6	6	6

Table 1: The Bit Allocation to the two sets of VQ codebooks.

Two sets of VQ codebooks were developed with their bit allocation shown in Table 1, requiring 4.7 kbps and 4.4 kbps respectively in transmission assuming 10 frames/sec. In each case the free syllable recognition accuracy (i.e., no knowledge about the lexicon and no language model) has been optimized as shown in Table 2. Initial/Final models were used in the Mandarin syllable recognition here, where Initial is the initial consonant for the syllable, while Final is everything after the Initial, including the vowel/diphthong part plus optional medials and nasal endings. It should be pointed out that the vector quantization produces quantization errors. Therefore if the syllable HMMs are trained with unquantized MFCC coefficients while the test MFCC coefficients are vector quantized, this was a mismatched condition just like noisy speech recognition by clean speech models. On the other hand, since the VQ codebooks are known to the server, the matched condition is achievable by simply using the vector quantized MFCC coefficients to train the syllable HMMs. In Table 2, the first row is the baseline test with syllable HMMs trained by unquantized MFCC coefficients and the test speech features also unquantized. The next two rows are for the mismatched condition. It can be found that with both VQ codebooks the syllable accuracy was

seriously degraded. The last two rows of Table2, are for the matched condition, i.e., the syllable HMMs are trained with vector quantized MFCC coefficients. It can be found that with both VQ codebooks and matched HMM training condition, only very slight accuracy degradation was observed.

VQ codebooks	Free Syllable Recognition Accuracy (%)			Bit Rate (kbps)
	Mix 2	Mix 4	Mix 8	
Baseline (without VQ)	54.02	60.79	67.75	
(A)8777766 (mis-matched)	40.06	45.51	51.02	4.7
(B)7766666 (mis-matched)	39.93	45.05	50.54	4.4
(A)8777766 (matched)	52.73	59.71	66.98	4.7
(B)7766666 (matched)	52.72	59.63	66.78	4.4

Table 2: Free Syllable Recognition Accuracy for different sets of VQ codebooks under matched/mismatched conditions.

The test database for the keyword spotting and verification includes utterances spoken by 10 speakers, each producing 100 utterances, with one key-word in each utterance, therefore a total of 1000 test utterances. The key-word set used in the test is the bank names in Taiwan, about 2400 in total. The key-word detection is based on forward search. We used the general filler model to detect the keywords through the forward search. The likelihood of the keyword is obtained by backtracking the Viterbi path.

The keyword spotting results are plotted as ROC curves as shown in Figures 1, 2, 3 and 4. Figure 1 shows the baseline results. The upper curve is the results when no vector quantization is performed, while the two lower curves are for the two vector quantization codebooks (A)8777766 and (B)7766666, without special efforts for keyword verification. Apparently the vector quantization errors actually degraded the keyword spotting performance, even if the degradation in syllable accuracy was already minimized. Also, it is cleat that lower bit rate gives lower performance. Figure 2 is the improvements achieved by keyword verification techniques for the codebook set (A)8777766. Here the lowest curve in Figure 2 is the baseline curve in the middle of Figure 1, while the two upper curves are obtained with 2-class(equation (2)) and N-best(equation (6)) verification approaches. We can see that the verification does offer very significant improvements, and the N-best approach proposed in this report is especially attractive. Figure 3 is similar to Figure 2 except it is for codebook set (B)7766666, so the lowest curve in Figure 3 is the lowest curve in Figure 1. Figure 4 finally compares all results using N-best verification(equation (6)), where the highest curve is for the case without vector quantization but with N-best verification applied (significantly higher than the baseline without vector quantization, as in Figure 1, also shown

here as the lowest

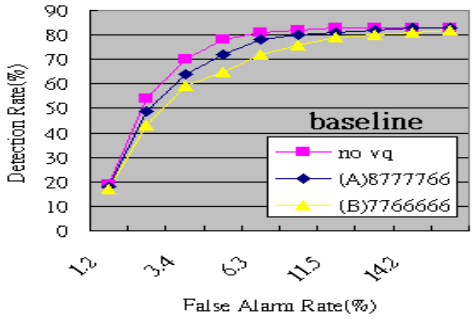

Figure 1: The baseline results: keyword spotting without special efforts on keyword verification.

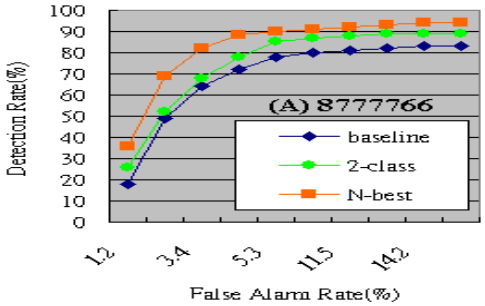

Figure 2: The results for codebook set (A)8777766: both 2-class and N-best verification techniques offers some improvements.

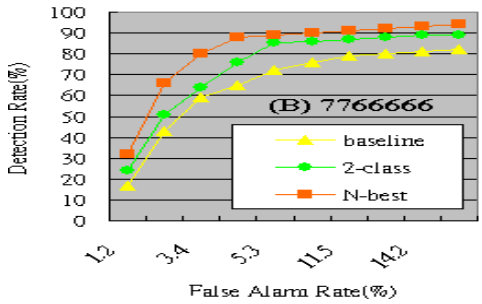

Figure 3: The results for codebook (B)7766666: both 2-class and N-best verification techniques offers some improvements.

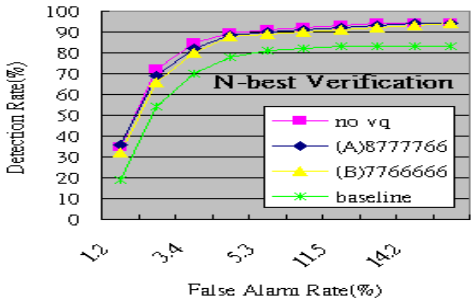

Figure 4: The best results obtained with N-best keyword verification compared to the baseline result without vector quantization.

curve), while the two curves in the middle are the best curves in Figures 2 and 3 for codebook sets (A)8777766 and (B)7766666. It can be found that with the proposed approach, not only the performance degradation caused by vector quantization is not a problem any longer, but even better performance becomes obtainable.

5. CONCLUSIONS AND FUTURE WORKS

In this study, we implemented vector quantization for Mandarin speech to be used in distributed speech recognition under wireless environment, and see the interesting results that Chinese keyword spotting performance can be significantly improved by keyword verification techniques with parameters optimized by MCE training. The studies in this report analyzed only the effects of quantization errors. Transmission errors and packet loss will be considered in the next step.

6. REFERENCES

[1] European Telecommunications Standards Institute, ``Speech Processing Transmission and Quality aspects (STQ); Distributed Speech Recognition; Front-end feature extraction algorithm; Compression algorithms", *ETSI Standard ES 201 108 v1.1.2, April 2000*.

[2] A. M. Kondoz, ``Digital Speech, Coding for low bit rate communication systems", *John Wiley & Sons*, 1999.

[3] Berlin Chen, Hsin-min Wang, Lee-feng Chien, and Lin-shan Lee, ``A*-Admissible Key-phrase Spotting with Sub-syllable Level Utterance Verification", in *Proceedings of International Conference on Spoken Language Processing, 1998*.

[4] Myound-Wan Koo, Chin-Hui Lee, and Biing-Hwang Juang, ``Speech Recognition and Utterance Verification Based on a Generalized Confidence Score", in *IEEE Transactions on Speech and Audio Processing, vol. 9, no. 8, Nov. 2001, pp. 821-30*.

[5] Frank Wessel, Ralf Schluter, Klaus Macherey, and Hermann Ney, ``Confidence Measures for Large Vocabulary Continuous Speech Recognition", in *IEEE Transactions on Speech and Audio Processing, vol.9, no. 3, Mar. 2001, pp. 288-98*.

[6] Tomoko Matsui, Frank K. Soong, and Biing-Hwang Juang: ``Verification of Multi-class Recognition Decision using Classification Approach", in *Workshop Automatic Speech Recognition Understanding(ASRU), 2001*.

[7] Wu Chou, ``Discriminant-Function-Based Minimum Recognition Error Rate Pattern-Recognition Approach to Speech Recognition", in *Proceedings of The IEEE, Aug. 2000, pp.1201-23*.

[8] Tatsuya Kawahara, Chin-Hui Lee, and Biing-Hwang Juang, ``Key-phrase Detection and Verification for Flexible Speech Understanding", in *Proceedings of International Conference on Spoken Language Processing, 1996*.