行政院國家科學委員會補助專題研究計畫成果報告

※※※※※※※※※※※※※※※※※※※※※※※※※

※　　　　　　　　　　　　　　　　　　　　　　　　　※

※　子計畫二：無線通訊環境下國語語音之分散式辨認　※

※　　　　　　　　　　　　　　　　　　　　　　　　　※

※※※※※※※※※※※※※※※※※※※※※※※※※

計畫類別：□個別型計畫　　■整合型計畫

計畫編號：NSC　89－2213－E－002－177－

執行期間：　2000 年　8 月 1 日至 2001 年 7 月 31 日

計畫主持人：李宇旼

共同主持人：

本成果報告包括以下應繳交之附件：

□赴國外出差或研習心得報告一份

□赴大陸地區出差或研習心得報告一份

□出席國際學術會議心得報告及發表之論文各一份

□國際合作研究計畫國外研究報告書一份

執行單位：

中　華　民　國　90　年　7　月　31　日

行政院國家科學委員會專題研究計畫成果報告

計畫編號：NSC 89-2213-E-002-177
執行期限：89 年 8 月 1 日至 90 年 7 月 31 日

主持人：李宇旼　國立台灣大學電信研究所
計畫參與人員：吳承晃 陳志豪 陳孝勇 黃品傑 高國浩 陳繹全
國立台灣大學電信研究所

一、中文摘要

人們致力於將整合語音技術到無線通訊網路已有一段時日。我們所關心的，是如何在低傳輸速率中提供可靠的辨識效能。在這篇報告中，我們首先會介紹分散式語音辨識的觀念，然後再討論向量量化以及通訊頻道中雜訊所帶給辨識效能的影響。

關鍵詞：分散式語音辨識，向量量化。

Abstract

There has been much effort to realize speech recognition technology in digital wireless network. All we concern about is to provide reliable performance at low transmission rate. In this paper, we first introduce some basic concepts about DSR (distributed speech recognition), and then we discuss the effects of vector quantization and channel noise on recognition performance.

Keywords: DSR (distributed speech recognition), vector quantization.

二、緣由與目的

People have dreamt of a universally accessible information database even before computer networking was invented. It is not until the 1990's have the technologies caught up to make such concept possible. The commercialization of the Internet and the invention of the World Wide Web (WWW) has prompted the creation of a universe of virtually unlimited, network-accessible information that range from private, secure databases to free, publicly available services. Such typical examples as digital libraries, virtual museums, distant learning, network entertainment, network banking, and electronic commerce have indicated that this global information network will soon become a physical embodiment of the whole human knowledge, and a complete integration of the global information activities. As the global information network evolves, efficient while user-friendly technologies for accessing the information infrastructure becomes increasingly important. Traditionally, the information infrastructure is accessed through a computer that is physically tied to a network. Commands are entered into the computer by the use of the keyboard and the mouse, while the retrieved information is displayed on the screen. Although reliable and economical, this mode of accessing the global information network will become increasingly inadequate as we progress into the 21st century.

The success of broadband wireless technologies will bring a new dimension into the way the above information is accessed or retrieved. Since full mobility support is possible, the user will no longer be constrained by the tether of the wire. The data rates of the future broadband wireless systems range from 384 kbps (third-generation cellular systems) to a few Mbps (broadband wireless access) to 10s of Mbps (wireless LAN), which are capable of carrying multi-media traffic. Therefore, with the successful deployment of broadband wireless communication systems, users should in principle be able to access the global network infrastructure at any time and from anywhere. However, as the size of the wireless terminal shrinks, traditional keyboard or keypad input and screen output becomes increasingly inconvenient. Therefore, an efficient, flexible, and user-friendly approach especially suitable for broadband personal wireless communicators is an important enabling factor for desired “anytime, anywhere” access to the information infrastructure. Speech is one of the most natural and user-friendly mechanisms for information access, and spoken language processing technologies, after diligent research efforts for decades, have matured to a point where many useful and commercially beneficial applications become feasible recently. Therefore, a good integration of advanced spoken language processing and broadband wireless technologies will be a key factor for the evolution of a wireless information community from the success of broadband wireless. DSR (Distributed Speech Recognition) is a client/server approach for speech recognition, where the client refers to the mobile terminal, while the server is a computer physically connected to the network infrastructure. In the following, We first introduce speech recognizer and then consider how to deploy the recognizer to the wireless communication systems.

Today’s state-of-the-art speech recognizers are based on statistical techniques, with hidden Markov modeling being the dominant approach. The typical components of a speech recognition and understanding system are the front-end processor, the decoder with its acoustic and language models, and the language

understanding component. The latter components extracts the meaning of a decoded word sequence and is an essential part of a natural language system. The remainder of this paragraph briefly reviews the frond-end and the decoder. The front-end processor typically performs a short-time Fourier analysis and extracts a sequence of observation vectors (or acoustic vectors) $X=[x_1,x_2,\ldots,x_T]$. Many choices exist for the acoustic vectors, but the cepstral coefficients have exhibited the best performance to date. The decoder is based on a communication theory view of the recognition problem, trying to extract the most likely sequence of words $W=[W_1, W_2,\ldots,W_N]$ given the set of acoustic vectors X. This can be done using Bayes' rule

$$\hat{W} = \arg\max_W P(W|X) = \arg\max_W \frac{P(W)P(X|W)}{P(X)}.$$

The probability P(W) of the word sequence W is obtained from the language model, whereas the acoustic model determines the probability P(X|W). In HMM-based recognizers, the probability of an observation sequence for a given word is obtained by building a finite-state model, possibly by concatenating models of the elementary speech sounds or phones. The state sequence $S=[S_1, S_2,..,S_T]$ is modeled as a Markov chain and is not observed. At each state S_t and time t an acoustic vector is observed based on the distribution $b_{St}=p(x_t \mid s_t)$, which is called output distribution. Because HMM's assume, for simplicity, that observations are independent of their neighbors, first- and second- order derivatives of cepstral coefficients are included in the acoustic vector x_t. If the acoustic vector generated by the front-end is passed directly to the acoustic model, then continuous-density distributions are used, with the multivariate-mixture Gaussians the most common choice

$$b_s(x_t) = \sum_{i=1}^{K} p(w_i|s)N(x_t; \mu_{si}, \Sigma_{si})$$

where $p(w_i|s)$ is the weight of the i-th mixture component in state s and $N(x; \boldsymbol{\mu}; \boldsymbol{\Sigma})$ is the multivariate Gaussian with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. In this work, we use continuous-density HMM's with mixture components that are shared across HMM states. Continuous density HMM's exhibit superior recognition performance over their discrete-density counterparts. From the above we can know that decoder with its acoustic model and language model has much more computation work than feature extraction does.

Then consider two basic observations. Firstly, the mobile client should not do too complicated work so that the kinds of client are not limited. Secondly, speech recognition needs only a small part of information that speech signal carries, and thus we can reduce transmission rate to save bandwidth if we only transmit the necessary information. Based on the two observations above, we can deploy the DSR system as follows. The mobile client may perform only the feature parameters extraction and compression. The compressed feature parameters are then transmitted over an error-protected wireless channel to the server, which is in charge of more complicated tasks of large-vocabulary speech recognition.

To advanced reduce transmission rate, we should do vector compression on the extracted feature vectors. But vector quantization will introduce quantization noise. Besides, different vector quantization will resulting different recognition performance: We can divide the acoustic vector into some sub-vectors with lower dimension and then do vector quantization on these sub-vector [1] . We find that different methods of grouping of sub-vectors lead to different results, and different size of codebook that assigned to the sub-vector also gives different results because usually the lower-dimension is more important. In our work, we will do some experiments to demonstrate those mentioned above. In addition, we know channel impairment will introduce errors to the transmitted featured vectors. In our work, we will design experiments to investigate the effect of channel noise brings.

三、結果與討論

In our experiment, we perform digit recognition and we use HMM as our model and MFCC as our feature set. Our training data and testing data are from Num-100, which is collected from 50 males and 50 females speakers through microphones. Training data are composed of 1000 single-digit utterances from 1000 files. We choose two sets of testing data : A set is composed of 1000 single-digit utterances from another 1000 files; B set is composed of 480 files with arbitrary length from single digit to seven digits. In the following we design two experiments to analyze the effect of vector quantization and the one of transmission error separately.

Firstly we investigate the effect of vector quantzation. We group two subsequent dimensions into a sub-vector and leave the energy component as an another sub-vector, and thus we obtain 7 ub-vectors. We only use testing data A set. We assign every sub-vector codebooks with different sizes. The result (Word Correction) is listed in Table.1. Note that word correction is defined as all words subtracted by substitutions and by deletions. In column 1 we list number of bits for sub-vector1, sub-vector2, … ,sub-vector7, etc. In column 2, 3, 4, 5, 6, we list number of mixtures per model. Results of recognition without VQ are also given in row 1. Our goal is to find the optimal bit-allocation for sub-vectors at fixed transmission rate. At bit rate= 3.5 kbps, we find that the optimal bit-allocation is 7665543. At this rate all results are similar, but if we take a close look at these data, it is obvious that we can obtain better result if we assign more bits for lower-dimension sub-vectors. At bit rate=3.4 kbps, the optimal bit-allocation is 6665443, and we also can obtain similar observation.

Then we investigate the effect of noise. Again

we group two subsequent dimensions into a sub-vector and leave the energy component as an another sub-vector, and thus we obtain 7 ub-vectors. We fix bit-allocation as 5555555. Testing set A and B are both used, and we give word correction and sentence correction (all list word in a sentence are matched) as results in Figure 1, 2, 3, 4. Note that all deep blue line are given as results for no VQ and error-free condition; all pink line are given as results for VQ and error-free condition; all yellow lines are given as results for VQ and BER=0.01; all blue lines are given as results for VQ and BER=0.001; all purple lines are given as results for VQ and BER=0.0001; all red lines are given as results for VQ and BER=0.00001. From these four figures we can know that if BER is 0.001, difference of recognition accuracy will not be higher than 4 %. If BER is below 0.001, the recognition accuracy is similar to whose BER = 0.001 without too much improvement Also we can observe that as number of mixtures increase the difference will be much lower.

From the two paragraphs above, we can conclude that: Under the same sub-vector structure and the bit rate is fixed, if we allocate more bits for lower dimension sub-vectors we can obtain better recognition performance; if we can reduce channel error to the degree of 10^{-3}, the recognition performance will be no longer dominated by channel noise.

Table.1 word correction obtained from different sub-vector grouping

Bits per sub-vector	1 mixture	2 mixtures	3 mixtures	4 mixtures	8 mixtures	Bit rate(kbps)
N0 VQ	94.01	95.60	97.20	98.00	98.20	
5555555	92.81	94.91	97.20	97.50	97.90	3.5
6666666	93.31	95.50	97.00	97.80	98.00	4.2
6664446	92.51	95.60	96.70	97.40	97.90	3.6
6664442	90.61	93.91	95.10	96.50	96.90	3.2
6655443	92.51	95.30	96.80	97.80	98.00	3.3
6664444	92.71	95.80	96.80	97.10	97.50	3.4
6655443	92.81	94.71	96.70	97.70	97.50	3.3
6655444	92.71	95.10	96.60	97.70	97.80	3.4
6665443	93.01	95.20	96.60	97.50	98.00	3.4
6666543	92.81	95.40	96.90	97.70	98.20	3.5
7665543	93.21	95.70	97.10	97.80	98.20	3.5
7765443	93.31	95.00	96.90	97.70	98.00	3.5

四、計劃成果自評

The results above give us some senses about the effect of quantization error and transmission error. We plan to lower transmission bit rate to about 2 kbps without degradation of recognition performance. Then we hope to find an adaptation rule so that under different transmission environment we can use suitable VQ codebooks to obtain better recognition performance.

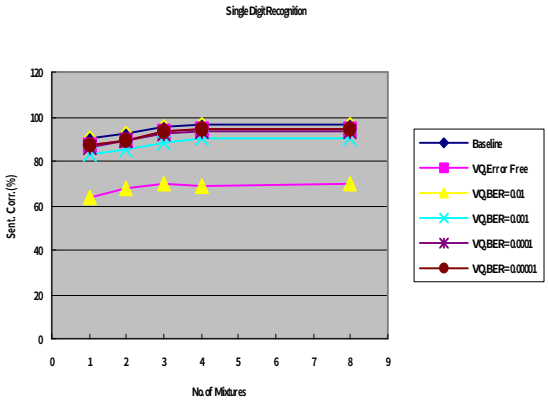

Figure.1 sentence correction for different BER (A set)

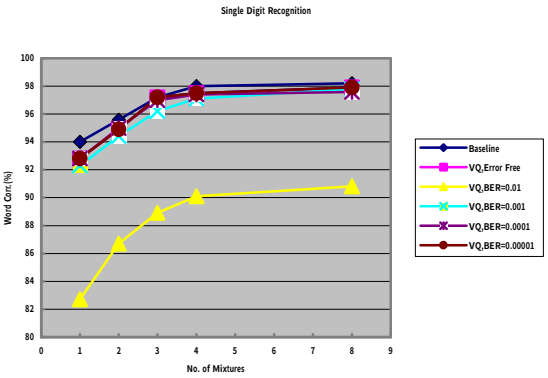

Figure 2 word correction for different BER (A set)

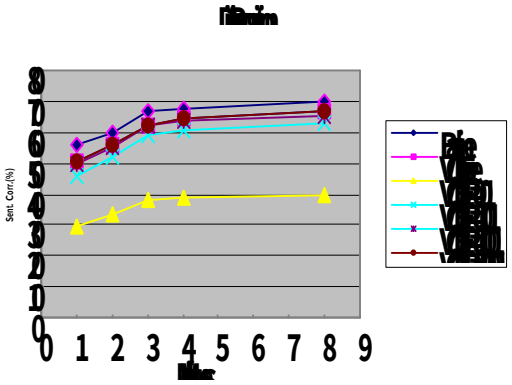

Figure 3 sentence correction for different BER (B set)

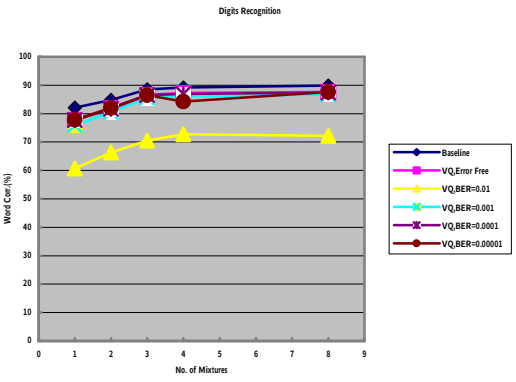

Figure 4 word correction for different BER (B set)

五、參考文獻

[1] V. Digalakis, L. Neumeyer, and M. Perakakis, ”Quantization of cepstral parameters for speech recognition over the world wide web,” *IEEE J. Select Areas Commun* ,vol. 17, 99. 82-90,Jan 1999.

[2] Lin-shan Lee, Yumin Lee, “Voice access of global Information for broadband wireless : technologies of today and challenges of tomorrow”