TANGERINE: A LARGE VOCABULARY MANDARIN DICTATION SYSTEM

Yuqing Gao, Hsiao-Wuen Hon, Zhiwei Lin, Gareth Loudon
S. Yogananthan and Baosheng Yuan

Apple-ISS Research Centre, Institute of System Sciences
National University of Singapore, Heng Mui Keng Terrace, Singapore 0511

ABSTRACT

The text input for non-alphabetic languages, such as Chinese, has been a decades-long problem. Chinese Dictation using large vocabulary speech recognition provides a convenient mode of text entry. In contrast to a character based Dictation system[5], a word-based Mandarin dictation system has been designed[3] (based on Apple's PlainTalk speech recognition technology[4]) for efficient entry of Chinese characters into a computer. In this paper new features and improvements to the Dictation system are presented. The new features and improvements have produced an overall reduction in recognition error of 50 - 80%. The vocabulary has also been increased from 5,000 words to over 11,000 words.

1. INTRODUCTION

For Chinese, high speed typing is achieved only by memorizing the mapping between each ideographic character into a sequence of keystrokes. With more than 6,000 Chinese characters, typing requires extensive training. Word-based Mandarin dictation using large vocabulary speech recognition technology has appeared as a very promising solution to Chinese text entry problem[3]. The dictation system we are presenting is a word-based, speaker-dependent speech recognition system, which uses HMM technology. In this paper, our emphasis has been on the new features and improvements to reduce the overall recognition error rate of the dictation system.

The new features and improvements we have investigated in this paper include the following aspects. MFCC analysis has been used to replace LPC derived Cepstral coefficients. Spectral subtraction has been applied for addition background noise reduction. Cepstral mean normalisation has been investigated as an approach to deal with the type of noise due to microphones and cables. HMM based tone classification has been used for tone recognition of polysyllabic words. The recognizer is HMM-based. Therefore, vocabulary independent training and the reduction of training data are the keys for a successful commercial system. This problem has been explored as well. Adaptive training of acoustic models has been used to track changes in a user's voice and in the environment. Enhanced acoustic models are being studied. A statistical language model is being used to overcome the problem of homophones in Chinese. In following sections, each problem is described independently.

2. ROBUST SPEECH REPRESENTATION

The previous Dictation system[3] used an LPC based speech representation. The new system uses Mel-scale frequency cepstral (MFCC) analysis[2] to describe the speech signal. The MFCC approach has reduced recognition errors by 17%. The MFCC provides a good representation of speech, but is not very robust to environmental and acoustical disturbances. This has led to the use of spectral subtraction and cepstral mean normalisation (CMN) methods to make the representation more robust.

There are two common types of noise disturbances. The first is additive. If the background noise is stationary the effect of the noise can be reduced by the use of spectral subtraction[1]. An estimate of the noise power spectrum is found and subtracted from the power spectrum of the speech signal during the MFCC analysis. The second type of noise is due to the microphone and cables. The frequency response characteristics of the microphone and cables introduces a frequency-dependent multiplicative factor on the spectrum of the speech signal. The effect of this noise can be reduced using cepstral mean normalisation[2]. One of the properties of the cepstral domain is that it converts a multiplicative effect in the frequency domain into an additional one. Therefore the frequency response characteristics of the microphone and cables is an additive effect in the cepstral domain. This effect can be reduced by finding a

0-7803-2431-5/95 $4.00

time averaged cepstral vector and subtracting it from every cepstral vector. The resultant cepstral vector will then be invariant to distortions introduced by the microphone and cables.

Spectral noise subtraction has no effect on accuracy at 20db SNR but has reduced errors by 25% at 10db SNR. Cepstral mean normalisation has reduced the recognition errors by an average of 34.5% for mismatch recording environments.

3. TONE CLASSIFICATION OF POLY-SYLLABIC WORDS

Mandarin is a tonal syllabic language and each syllable is assigned one of 5 tones. Syllables of the same phonetic structures but with different tones usually convey different meaning.

Although the tonality of a Mandarin syllable is mainly characterised by pitch contour, we found the derivative of pitch signal is also essential for distinguishing different tones. The previous tone recognition approach[3] was a post process to the sub-syllable HMM recogniser. The method found the underlying pitch of an utterance and performed a template matching procedure against four templates representing the four tones in Mandarin. This approach achieved high tone recognition accuracy for monosyllabic words but was not extensible to poly-syllabic words. This is because of the effect co-articulation has on the tonal characteristics and the requirement of pre-segmentation. To overcome this limitation, the pitch signal of the utterance and its derivative were used as an additional input to the HMM recogniser rather than as input to a separate process[7]. A new codebook was used to describe the pitch characteristics. The integration of the pitch information into the HMM recogniser has two main advantages:

(1) The approach handles context dependent tonal characteristics.

(2) It is consistent with the HMM frame-synchronised search architecture, avoiding the need for pre-segmentation.

The result is a reduction in recognition error of 50% for poly-syllabic and mono-syllabic words.

4. VOCABULARY INDEPENDENT TRAINING AND TRAINING DATA REDUCTION

Although an HMM based algorithm usually needs a large quantity of training data to achieve a high accuracy, it is very important for a speaker-dependent speech recognition system to keep the amount of training speech to a minimum while maintaining the required performance. Therefore, it is impossible for the system to use word models or syllable models and keep the training data concise[6]. Therefore vocabulary independent, sub-syllable models[3] have to be applied. In our system 250 context-dependent sub-syllable phonetic units have been identified for acoustic modelling. For a hidden Markov model that requires 20-30 speech samples for training, the whole system would need a total number of 5,000 to 7,500 phonetic units of Mandarin for HMMs training. This translates to 2,500 to 3,750 syllables for training. However, it is very difficult to form a meanful training script containing say 3000 syllables that has an equal number of sub-syllable phonetic units. Therefore the actual number of words selected for the training script is slightly larger. The following formula can be used to get a phonetically balanced training script by selection of the words by score:

$$C(V) = [\sum_{j=1}^{J}(1 - \frac{N(M_j)}{N_0})]W(V)$$

where, $C(V)$ is the score of the examined word V, M_j, $j = 0, 1, ...J$ is the jth phonetic unit in word V, $N(M_j)$ is the count of the unit M_j existing in the target script, N_0 is the target count for a unit, $W(V)$ is a weight for word V, which is determined by the using frequency of word V.

A speech database of 10,000 utterances of polysyllabic words and 3,684 utterances of monosyllabic words were used to form the phonetically balanced training script. From this speech database 2,400 polysyllabic words and 2,000 monosyllabic words have been selected using the above formula. The test set consists of 10,000 utterances of polysyllabic words and 1,228 of monosyllabic words. The test results listed in Table 1 below shows that when the training script has a good phonetic balance the training data can be reduced by more than three times without significantly affecting recognition performance.

In Table 1, Set 1 of training data is the original speech database, which has 10,000 polysyllabic words and 3,684 monosyllabic words. Set 2 and Set 3 are subsets of Set 1 selected by the above formula. Set 2 includes 3,000 polysyllabic words and 2,300 monosyllabic words. Set 3 includes 2,400 polysyllabic words and 2,000 monosyllabic words. On the average, each phonetic unit has 28 and 20 speech samples for Set 2 and Set 3, respectively. MS, PS and MPS in Table 1 mean the monosyllabic word, the polysyllabic word and the mix of mono- and poly- syllabic words, respectively.

Training data	Top 1(%)			Top 5 MPS(%)	Data reduction
	MS	PS	MPS		
Set 1	88.4	95.4	94	99.2	-
Set 2	88.3	94	92.9	99	2.8 times
Set 3	87.8	92.9	91.9	98.8	3.5 times

Table 1. Results of training data reduction

5. ADAPTIVE TRAINING

Although the system is speaker-dependent, only a small amount of training data is used. More training data will always be helpful for improving recognition performance. Furthermore, acoustic models should be updated gradually to track changes in a user's voice (such as speed, stress, loudness), and in the environmental changes. A solution to collecting more data and updating the acoustic models is to incrementally adapt HMMs. The adaptation is performed as an off-line batch process. The new training data used for the adaptation is speech data correctly recognised and stored during a previous dictation session when the user was using the system entrying his text. The adaptation is for the HMM output distributions only. The phonetic balance on the new training data cannot be guaranteed because it depends on the user's dictation text content. The adaptation model parameters may therefore suffer from sparseness. To overcome this problem, two types of adaptation models are used for adaptation training: context-independent and context-dependent. The deleted-interpolation algorithm is used to smooth context-dependent models with context-independent models. The resultant context-dependent adaptation model is then combined with the original model according to each one's count contribution. Table 2 below shows the results of the adaptation procedure. VQ codebooks were not retrained during adaptation.

Training set	Error reduction
Baseline: 2000 words	-
Plus adaptation data:	
1500 words	8.8%
3000 words	18.9%
6000 words	27.2%

Table 2. Results of adaptive training

6. ENHANCED SUB-SYLLABLE ACOUSTIC TRAINING AND RE-SCORING

The use of one fixed type of HMM structure for the 27 initial and 38 final sub-syllable acoustic models is done for simplicity. In fact, initials are generally shorter in duration and finals are longer. They can also be classified into groups according to their pronunciations and durations as follow.

AH, EH, OH, WH, YH, UUH;
B, D, G;
P, T, K;
G, K, H;
M, N, L;
J, Q, X;
Z, C, S;
ZH, CH, SH;
A, E, I, O, U, UU;
AI, AO, EI, ER, IA, IAO, IE, IU, IZ, OU, UA, UAI, UI, UO, UUE;
AN, ANG, EN, ENG, IAN, IANG, IN, ING, IONG, ONG, UAN, UANG, UEN, UENG, UN, UUAN, UUN.

Enhanced acoustic models are being studied where different structures of HMMs are being used to model different classes of phonetic units. Some preliminary experiments have been performed where one HMM structure is used for all the initials and another HMM structure, with more states, is used for all the finals. Initial results show that an error reduction of 3.5% can be achieved with this enhanced modeling procedure.

7. STATISTICAL LANGUAGE MODEL BASED HOMOPHONE PROCESSING

A statistical language model (SLM) is being used to overcome the problem of homophones in Chinese. It is a well known fact that the Chinese language is abundant with cases of homophones - when a syllable can map itself to more than one Chinese ideograph. As an example, the syllable *bang1* can be represented as 邦(state, county or kingdom),帮(assistance),梆(wooden bat), and 浜(waterfront). Such a problem arises because there are 1229 syllables in the language but there are about 6000 ideographs to which they can map onto.

While acoustic analysis can recognize the syllables of a spoken utterance with rather high accuracy, it has no means of deciding on an ideograph should a case of homophony arise. In current Mandarin speech systems, the user is made to choose the correct ideograph representation from a list of possible ones. This is, both distracting as well as, it is deviant from the hands free, eyes free interaction environment that dictation systems are supposed to provide. In our isolated word Chinese dictation system, we have attempted to reduce the amount of system initiated interaction between the user and itself by using a Statistical Language Model (SLM) to aid in the homophone problem by predicting the correct ideograph representation. Our language model was applied towards rescoring acoustic scores of a test data set consisting of running text. A study of the results for the monosyllabic words in the text,

showed that the language model helped by reducing errors by about 50%.

Currently a few different types of SLMs are being studied for inclusion into our system. Apart from their augmentation capabilities, practical details of implementation are also being investigated. The SLMs we are presently using are, a word unigram working either in isolation or with complimentary aid from a standard word bigram, or a tail character word bigram, or a transition character word bigram. The data for these SLMs was culled from about a 60 Mbyte corpus of Chinese text drawn from various news oriented sources in China. At the word level, the lexicon consists of about 57K tokens, while at the character level, it consists of about 5.8K tokens. Before generation of the SLMs, the given corpus, after cleaning, had to be segmented (word boundary marker inclusion). This was done using a relaxation segmentation algorithm with[8] the additional aid of a dynamic programming based secondary processor. The SLMs are generated in the standard way, while the smoothing of scores was done using a linear-discounting algorithm.

In our system, the language model is applied as a post-processing process to the acoustic analysis. The n-best list together with their relevant acoustic scores for an utterance are input into the SLM rescoring module. From the n-best list, all the ideographs mapping onto the n-best list are extracted. The ideographic list is re-ranked according to their combined scores, which are obtained from the acoustic scores and all included language model scores:

$$P = W_1 * P_a + W_2 * P_u + W_3 * P_b$$

where, $W_i, i = 1, 2, 3$ are the weights for acoustic, unigram and bigram score.

SLMs	Error correction%	
	Pinyin	Hanzi
Unigram	50.0	39.8
Plus Character Bigram	51.2	45.1
Plus Tansition character Bigram	51.8	45.8
Plus Tail Character Bigram	52.8	46.2
Plus Word Bigram	56.4	48.3

Table 3. Results of SLM rescoring

8. SUMMARY

All the above isolated experimental results strongly indicated that the overall performance of the system would significantly improve. Experiments on the integration of all these features and further improvements are still going on. According to the preliminary experiment result, the integration of these new features has produced 50-80% error reduction for 5,000 words baseline system. When the system vocabulary increased to over 11,000 words, the performance remains the same as 5,000 words baseline system. Among these 11,000 words, there are more than 10% monosyllablic words, increasing the ambiguity in recognition.

9. REFERENCES

[1] Atal, B.S., "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification", The Journal of the Acoustical Society of America, pp. 1304-1312, Vol 55, June 1974.

[2] Davis, S.B. and Mermelstein, P., "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", IEEE Transactions on Acoustics, Speech and Signal Processing, pp. 357-366, Vol. 28, August 1980.

[3] Hon, H.W. et al., "Towards Large Vocabulary Mandarin Chinese Speech Recognition", The Proceedings of ICASSP, pp. 545-548, 1994.

[4] Lee, K.F., "The Conversational Computer: An Apple Perspective", The Proceedings of EUROSPEECH, pp. 1377-1384, Sept. 1993.

[5] Lin-shan Lee et al., "Golden Mandarin (II) - An Improved Single-Chip Real-Time Mandarin Dictation Machine for Chinese Language with Very Large Vocabulary", The Proceedings of ICASSP, pp. 503 - 506, April, 1993

[6] Wang Hsin-Min, Chang Yuen-Chen and Lee Lin-Shan, "Automatic Selection of Chinese Syllable-Balanced Sentences from Chinese Text Corpus", pp. 195. ROCLING-IV, 1993

[7] Yang, W., Lee, J., Chang, Y. Wang, H., "Hidden Markov Model for Mandarin Lexical Tone Recognition", IEEE Transactions on Acoustics, Speech, and Signal Processing. Vol 36, pp 988-992, July, 1988.

[8] Fan. C.K., Tsai, W.H., "Automatic Word Identification in Chinese Sentences by the Relaxation Technique", Computer Proceedings of Chinese and Oriental Languages. Vol. 4, No. 1, pp. 33-56, 1988.