High Accuracy Chinese Speech Recognition Approach with Chinese Input Technology for Telecommunication Use

York Chung-Ho YANG, June-Jei KUO

Matsushita Electric Institute of Technology (Taipei) Co., Ltd., Taipei, Taiwan, ROC.
E-mail: {york, kaku}@mitt.com.tw
FAX: + 886 2 27556005

ABSTRACT

Phoneme-oriented input with synchronised auto-revision to pictographic nature of written Chinese and Chinese speech recognition are two subjects not often brought together in the same article nor even the same proposition. This paper explores the growing relations between these two entities and, in particular, investigates what is found in integration the Chinese phonetic input with its auto-revision methods (Kuo; 1986, 1987, 1995, 1996) and Chinese isolated word, continuous speech recognition for portable device such as mobile telephone. Chinese phonetic input with a synchronised auto-revision approach integrates with a small size, high recognition rate Chinese speech recognition kernel for DSP single chip application will be introduced in this paper. Chinese phrase taxonomy has been defined and the definition is ready to be obtained from the system's dictionary.

1. INTRODUCTION

The system we proposed here, we expect that it could input Chinese characters by only 11 keys that could show the traditional Chinese phonetic Zhu-Yin or Pin-Yin input symbols and Chinese characters into a portable device. As we knew, the phoneme-based approaches have been proposed and used for a long while in Chinese society. This paper, we will discuss a robust Chinese real-time synchronised auto-revision input system. To realize the speech recognition technology in real applications, speech recognizer must be robust to noisy environments and spot intended words from background noise and unintended utterances. Furthermore speech recognizer must retain high quality performance on portable devices. That is, small size in its executable code. Usually standard patterns for speech recognition are made by statistically processing speech data of speakers. There are several matching methods: for example, a method using the statistical distances measure, and a method applying models such as the neural net and HMM. Especially, numbers of successful HMM are reported using the continuous mixture Gaussian density models. With these methods, spectral parameters are used in speech recognition as a feature parameter and an enormous number of speakers are generally required for training. If the standard patterns for speaker independent speech recognition can be produced from a small number of speakers, manpower and computer power are saved and speech recognition technique can be easily handled in various applications. For the purpose mentioned above, we propose a speech recognition method using the similarity vectors as feature parameters for Chinese small scale vocabulary, isolated word, and continuous speech recognition. In this method, word templates trained with a small number of speakers, yield high recognition rates in speaker-independent recognition. We will also report the high performance by using the word templates that were made by concatenating sub-word units such as CV and VC, extracted from a small number of speakers. The integration of Chinese input method and Chinese speech recognition technology into telecommunication system can provide better human machine interface especially in portable device, such as auto dialing or Chinese message delivery on mobile phone. Also the integrated system can provide a good paradigm for the direction of Chinese CTI's development.

2. CHINESE INPUT ALGORITHMS

It is necessary to realise that Chinese text and massive Chinese information should be computerised in order to the convenience of data access, information retrieval and extraction, index searching and so on. How to improve the quality of computerised text will be a significant topic, however. Face various computational Chinese issues, this paper attempt to propose an intelligent auto-revision Chinese input system for portable telecommunication device. An auto-revision system for Chinese system, a crucial idea is comprehensive that lexicon knowledge plays an important role. Lexicon knowledge refers to the component of a fundamental NLP system that contains semantics, grammatical rules and so on. In Chinese, however, lexicon knowledge has to add the factors of pronunciation, i.e. phonological variance. The most important feature in Chinese phonological structure is that Chinese contains a huge set of heteronym distributing in many Chinese characters. In other words, Chinese holds an enormous number of homonym characters. Some Chinese characters can be pronounced to several different pronunciations and of that which kind of pronunciation can be shown that depends on its role in Chinese text. Because each Chinese character merely includes a monosyllable and a tongue, consequently, any isolated

Chinese character is extremely ambiguous. Here the examples are shown below:

*畜生: **chu4** sheng1 :beast, domestic animal*

*畜牧: **xu4** mu4: raise livestock*

*安步當**車**: an1 bu4 dang1 **jv1: steady and leisurely walking***

*車水馬龍: **che1** shuei3 ma3 lueng2: numerous cars go and around*

where the isolated characters 畜 and 車 can be pronounced as **chu4** and **xu4**, **jv1** and **che1** respectively. Secondly, in traditional Zhu-Yin system, some phonemes are very confusing to almost all Chinese, such as ㄓ (zh) and ㄗ (z), ㄔ (ch) and ㄘ (c), ㄙ (s) and ㄕ (sh), ㄣ (en) and ㄥ (eng) and so on. As the result, it frequently happens that the computer screen shows the Chinese character, which is not expected by users. The following examples illustrate one correct phrase and two valid but unsuitable phrases for the whole sentence.

聲(ㄕㄥ)色(ㄙㄜˋ)sheng1 se4 -> voluptuous

深(ㄕㄣ)色(ㄙㄜˋ)shen1 se4(valid but unsuitable)->deep colour

申(ㄕㄥ)設(ㄕㄜˋ)sheng1 she4(valid but unsuitable)->application

場所很容易成爲治安的死角-> a place where is easy to be a weakness point of peace.

In any monosyllable Chinese characters, the factor of tongue is extremely crucial and variances, thus even though native Chinese can not distinguish or realize it well. Therefore, from the examples above, we know that Chinese phrases are ambiguous because not only their homophones but also Chinese has two-character words, three-character words, and four-character words which may have many possible homophones. Those examples above show possible miswriting characters in every two-character word. The same errors also can be made in three-character or four-character words, or in a common sentence.

To solve the problem of traditional Chinese text, we propose that Chinese phonetic information and mask approach can deal with the input errors. The mask algorithm was firstly proposed by Kuo (1986, 1987, 1995, 1996, 1997) that describes two important parts. One is for phoneme mask, the other is character mask. In our definition, each Chinese phoneme includes 11 bits and the rule of their representation can be viewed below.

Table 1 Bit Mask for Chinese Phonemes

ㄓ->	1	0	1	1	0						
ㄗ->	1	0	1	1	1						
ㄧ->						1	0				
ㄩ->						0	0				
ㄛ->								0	1	0	0
ㄜ->								0	1	0	1

From the information above, if we would like to have the phonemes "ㄒㄩㄝ 2 ㄒㄧㄠ 4", but a mistyped occurred and to be "ㄒㄧㄝ 2 ㄒㄧㄠ 4." The mask technology now is used, then. According to the algorithm, the mistyped phoneme still can be automatically corrected by matching the counterpart Chinese word.

On the other hand, the character mask technology is that Chinese words refer to the candidate words of dictionary and can extract right character. For instance, there are continuous syllable characters A, B. The candidate model should be:

* AB, A*B, AB *, *A, A*; *B, B*

3. CHINESE SPEECH RECOGNITION

There are several matching methods: for example, a method using the statistical distances measure, and a method applying the neural net models, and Hidden Markov Model (HMM). Especially, numbers of successful HMM are reported using the continuous mixture Gaussian density models. With these methods, spectral parameters are used in speech recognition as a feature parameter and an enormous number of speakers are generally required for training. If the standard patterns for speaker independent speech recognition can be produced from a small number of speakers, the size of computation will be much smaller than usual. Therefore, human power and computation are saved and speech recognition technique can be easily handled to various applications. For the purpose mentioned above, we proposed our approach using the similarity vectors as feature parameters. In this method, word templates trained with small number of speaker yield high recognition rates in speaker-independent recognition. Our approach is for developing a high accuracy speaker-independent Chinese speech recognition system using the similarity vectors as feature parameters. An empirical result of word recognition rate is 97.5% with 106 cities cover Taiwan based on noisy environment. Our approach of accuracy rate in Chinese speech recognition has much higher than conventional methods. The present approach overcomes the deficiency and limitations of the prior art with a system and method for recognizing Mandarin Chinese speech with small number of training speakers. The present approach advantageously implements in a size-intensive device when determining the Initial and Final of a syllable to identify the phonetic information of a Chinese word.

Our approach begins with a user creating a speech signal to accomplish a given task. In the second step, the spoken output is first recognized in that the speech signal is decoded into a series of phonemes that are meaningful according to the phoneme templates. The acoustic analysis part analyses speech inputs and the extracted LPC (Linear Predictive Coding) cepstrum coefficients and delta power. The extracted parameters are matched with many kinds of phoneme templates, and static phoneme similarity and the first order regression coefficients of phoneme similarity are calculated in the similarity calculation part. After that, the time sequence of those number of phoneme templates to define a dimensional similarity coefficient vectors and regression coefficient vectors can be obtained. In the similarity calculation part, mahalanobis' distance algorithm is

employed for distance measure, where covariance matrixes for all of the phonemes are assumed to be the same. The meaning of the recognized words is obtained by the post processor that uses a dynamic programming to satisfy the real world that the input sounded and what it has previously recognized by phoneme similarity calculation. Consequently, the post processing make a decision according to the previous phoneme result that reduces the complexity of all the recognition model. Finally, the recognition system responds to the user in the form of a voice output, or equivalently, in the form of the requested action being performed, with the user being prompted for more input.

The LPC method has been used in a large number of recognizers for a long time. In particular, the basic idea behind the LPC model is that a given speech sample at time *n*, *S(in)*, in the preemphasis box, can be approximated as a linear combination of the past *p* speech samples, such that

$$S'(n) \cong a_1 S(n-1) + a_2 S(n-2) + \ldots + a_p S(n-p)$$

where the coefficients $a_1, a_2, \ldots\ldots, a_p$ are assumed constant over the speech analysis frame. We define the value $a_1, a_2, \ldots\ldots, a_p$ as 0.95. In the step of the Frame Blocking, the previously dealing of the preemphasized speech signal, *S'(n)*, is blocked into frames of *N* samples, with adjacent frames being separated by *M* samples. Assume we denote the l^{th} frame of speech by $x_l(n)$, and there are *L* frames within the entire speech signal, then

$$x_l(n) = S'(Ml + n), \qquad n = 0, 1, \ldots\ldots, N\text{-}1, \quad l = 0,1,\ldots,L-1$$

The values for N and M are 300 and 100 corresponding to the sampling rate of the speech are 8kHz.

After the LPC analysis coefficients have been done, LPC Parameter Conversion to Cepstral Coefficients processing is going to be dealt. This very important LPC parameter set, which can be derivated directly from the LPC coefficient set, is the LPC cepstral coefficients, c(*m*). The recursion used is:

$$C_0 = \ln \delta^2$$

$$C_m = a_m \sum_{k=1}^{m-1} \left(\frac{k}{m} \right) C_K a_{m-k}, \quad 1 \le m \le p$$

$$C_m = \sum_{k=1}^{m-1} (k/m) C_k a_{m-k}, \qquad m > p$$

Where δ^2 is the gain term in the LPC model. So until the description above, we have got the input vector *C* composed of LPC cepstrum coefficients and delta power in many frames.

In the similarity calculation part, we employ the simplified Mahalanobis's distance for distance measure, where covariance matrixes for all the phonemes are assumed to be identical. Input vector *c* is composed of LPC cepstrum coefficients, delta power in 10 frames. The input vector *c* is expressed as:

$$c = (v^1, c_0^1, c_1^1, \ldots, v^{10}, \ldots, c_{13}^{10})^t$$

where c_i^K denotes the *i*-th LPC cepstrum coefficient of the *k*-th frame and v^k denotes delta power of the *k*-th frame.

The phoneme similarity between input vector *c* and phoneme template (phoneme *p*) is calculated as

$$L_p = a_p \cdot c - b_p$$
$$a_p = 2\Sigma^{-1} \cdot \mu_p$$
$$b_p = \mu_p \cdot \Sigma^{-1} \cdot \mu_p$$

where μ_p is a mean vector of phoneme *p*, and Σ is the covariance matrix.

After the static phoneme similarities are obtained, regression coefficients of the phoneme similarities are computed using static phoneme similarities over 50 msec. The word templates are produced by concatenating sub-word units such as CV and VC trained from a few speakers' speech. Especially, in the similarity calculation portion, it includes phoneme-templates that consist a Chinese Initial field and a Chinese Final one. Accordingly, the similarity parameter can be obtained by the calculation of *s(i, j)*, which is the score function to calculate the partial similarity.

$$s(i,j) = w \frac{d^i \cdot e^j}{\left|d^i\right| \cdot \left|e^j\right|} + (1-w) \frac{\Delta d^i \cdot \Delta e^j}{\left|\Delta d^i\right| \cdot \left|\Delta e^j\right|}$$

where d^i denotes a similarity vector in the *i*-th frame of input, e^j denotes a similarity vector in the *j*-th frame of reference, and Δd and Δe^j are the respective regression coefficient vectors, and '*w*' is the mixing ratio between scores from the similarity vector and its regression coefficient vector. The trajectories of the similarity are regression coefficients are averaged for each sub-word unit and stored in a sub-word dictionary. In the matching portion, we employ the most widely used technology that is well known as "Dynamic time Warping (DTW)" for our word template recognition processing. DTW is fundamentally a feature-matching scheme that inherently accomplishes "time alignment" of the sets of reference and test features through a DP procedure. By time alignment we mean the process by which temporal regions of the test utterance are matched with appropriate regions of the reference utterance. The need for time alignment arises not only because different utterances of the same word will generally be of different duration, but also because phonemes within words will also be of different duration across utterances. The Dynamic Programming for word matching with word templates algorithms are shown as:

$D = \sum_{k=1}^{K} d_N(i_k, j_k)$, $t(i_k)$ matches with $r(j_k)$, for

k =1, 2, ..., K

is the path (i_k, j_k), for *k=1, 2, ..., K*

the accumulated distance is, for example, *g(i, j)*

$$g(i,j) = \max \begin{matrix} g(i\text{-}2, j\text{-}1) + s(i,j) \\ g(i\text{-}1, j\text{-}1) + s(i,j) \\ g(i\text{-}1, j\text{-}2) + s(i, j\text{-}1) + s(i,j) \end{matrix}$$

In the empirical result, based on 106 city names cover

Taiwan, the table as following shows the accuracy of traditional LPC cepstrum coefficient recognition rate.

Table 2 Traditional LPC Accuracy Rate

Precision of Feature Parameters	32bit	8bit
LPC Cepstrum Coefficients Recognition Rate (%)	84.3	74.1

On the other hand, based on the same experimental data, the empirical result of our approach below shows the accuracy rate has been much improved by our algorithm.

Table 3 Similarity Vector Accuracy Rate

Precision of Feature Parameters	32bit	8bit
Similarity Vector Recognition Rate (%)	97.5	97.5

4. COMBINED CHINESE INPUT AND SPEECH RECOGNITION TECHNOLOGY TO PORTABLE DEVICE

These two technologies have been combined together for telecommunication use. The diagram below shows a model of our system in a mobile device. We designed a useful interface for Chinese input with just 11 buttons. For example, if we press songxia, using ←and → keys for aid, it will automatically show Chinese characters:"松下" based on our algorithm and system's phrase taxonomy. After that, phone number can be registered. Because the system embeds a speech recognition kernel, we can use this function for auto dialing, for example.

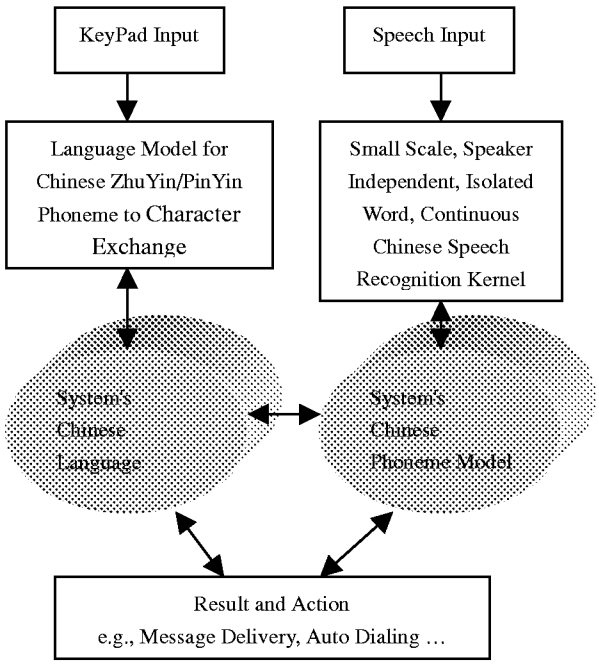

Figure 1 System Diagram of a Mobile Device Paradigm

5. CONCLUSION

To sum up, some points of view can be described. Firstly, our Chinese input system significantly improves some traditional Chinese text error-correction approaches. Secondly, there is no necessary or demand in special knowledge database for aid. Thus it cause huge cost down not only in human works but also in money in maintain or collection for a knowledge base. Kuo has been remarkably successful in pulling together in the elements of phonetic-oriented modification and a foray of model-theoretic semantics into its synchronised words error-correction. Thirdly, with an embedded Chinese speech recognition kernel to mobile phone application is a big advantage for human-machine interface. This system can be applied to short message delivery because of its remarkable Chinese input technology and it also can be used for auto dialing because of its accuracy Chinese speech recognition technology.

BIBLIOGRAPHY

[1] Chang, C.H., "A pilot study on automatic Chinese spelling error correction," *Communications of COLIPS*, Vol. 4, No. 2, 1994, pp.143-149.

[2] Hosimi M., M. Miyada, S. Hiraoka and K. Niyada, "Speaker-Independent Recognition Method Using Training Speech from a Small Number of Speaker", ICASSP-92, 1992, ppI-469.

[3] Hosimi M., M. Miyada, S. Hiraoka and K. Niyada, "Speaker-Independent Speech Recognition Method Using a few Person's Utterance as Model of Dynamic Characteristic of the Word Speech", IEICE Technical Report SP91-20, 1991.

[4] Kuo, J-J, Jou, J-H, Hsien, M-S, Maehara, F, "The Development of New Chinese Input Method – Chinese Word-String Input System," *Proceeding of International Computer Symposium*, 1986, Tainan, Taiwan.

[5] Kuo, J-J, "Chinese Document Revision System using Phonetic Information," Unpublished MITT Technical Report, 1996, Taipei, Taiwan.

[6] Leung, C.H., Kan, W-K, "Difficulties in Chinese Typing Error Detection and Ways to the Solution," *COMPUTER PROCESSING OF CHINESE & ORIENTAL LANGUAGES*, Vol. 10, No. 1, 1996, pp. 97-113.

[7] Park, J.K., Lee, S.H., Kim, J.H., "An error correction algorithm for Hangul Text Recognition," *COMPUTER PROCESSING OF CHINESE & ORIENTAL LANGUAGES*, Vol. 5, No. 2, 1991, pp. 183-192.