A CONTINUOUS PUTONGHUA RECOGNIZER

Pak-Kwong WONG and Chorkin CHAN

Department of Computer Science
The University of Hong Kong
Pokfulam Road, Hong Kong.
pkwong@cs.hku.hk and cchan@cs.hku.hk

ABSTRACT

A multi-speaker continuous Putonghua recognizer has been developed composing of 20 speaker-dependent recognizers as sub-systems. Each sub-system is a network of hidden Markov models modeling triphones as the fundamental speech units. Over 3GB of speech data have been collected for training from twenty native Putonghua speakers reading carefully designed texts trying to include all phone-to-phone transitions in Putonghua. A Viterbi path search yields the best speech unit sequence over the HMMnet for each unknown input utterance which is then passed down to a language model for post-processing. The most suitable word sequence is determined by means of the bigram statistics of 470 word classes covering a vocabulary of over 80,000 words. An enrollment process is required for each new user to select the most suitable speaker-dependent system among the 20 sub-systems according to their recognition performances on a small quantity of speech data collected from the user.

1. INTRODUCTION

Due to the non-alphabetic nature of the Chinese language, computer entry of Chinese information by means of a keyboard is naturally inconvenient. The urge for a Chinese dictation machine is therefore understandable. Six Putonghua recognizers of large-vocabularies for that purpose have been announced in the past [1, 2, 3, 4, 5, 6]. All six systems, developed by two research teams respectively, are speaker-dependent. The first three are recognizers of isolated syllables, the fourth one recognizes isolated words and the remaining two recognize connected speech. This paper reports the development of a multi-speaker, very large vocabulary, connected Putonghua recognizer at the University of Hong Kong.

2. SPEECH DATABASE FOR SYSTEM TRAINING

A Putonghua corpus of isolated syllables, words, digit strings and connected speech has been constructed in 1993 at the University of Hong Kong. A total of 20 native Putonghua speakers are employed to read prompting messages displayed on a monitor screen.

All recordings are done in a quiet office with a microphone (National Cardioid Dynamic Microphone WM-333N IMP600Ω) and a Sound Blaster 16 ASP A/D-D/A card sampling at 16KHz. The text contents of this database include isolated syllables, words, digit strings, rhymed syllables, continuous speech, and retroflexed ending words. The corpus of waveforms thus captured exceeds 3GB in total.

Despite the abundance of continuous utterances read by each speaker, certain phone-to-phone transitions between syllables are still missing. Five of the speakers read extra words to make sure all transitions of the last phone of a syllable to the first phone (initial) of a syllable are covered. Furthermore, two speakers read an additional subset of words each retroflexed at the end.

Utterance end points are determined automatically according to the energy and zero crossing rate profiles of the utterance. Each utterance is preceded and followed by at least 15 mseconds of silence. The speech data produced by a speaker are used to train a speaker specific sub-system, which, in turn, is used to phonetically label each utterance from that speaker according to the Viterbi path over a specific state sequence compatible to the utterance.

The twenty speakers, ten females and ten males, who generate this speech database, are aged between 20 and 39. They all are native Putonghua speakers except one who is a Hong Kong resident girl speaking with a mild accent.

3. SHORT-TIMED ACOUSTIC FEATURES AND RECOGNIZER ARCHITECTURE

Frames of 20 mseconds are Hamming windowed and 3/4 overlapped with its immediate neighbors. From each frame, 12 MFCCs, 12 Δ MFCCs, the energy and the Δ energy are derived. Δ features of a frame are computed from frames 15 mseconds before and after it. These features are divided into 3 streams with a codebook of size 512 for the MFCC vectors, a codebook of 256 codewords for the Δ MFCC vectors, and a codebook of 128 for the vectors of energy and Δ energy. Each codeword is a Gaussian density function with a diagonal covariance matrix to serve as a mixture component shared by all state densities:

0-7803-4137-6/97/$10.00©1977 IEEE

$$s_m(\mathbf{O_t}) = [\sum_{i=1}^{512} c_i^{(MFCC)} N^{(MFCC)}(\mu_\mathbf{i}, \Sigma_\mathbf{i}; \mathbf{O_t})] \times [\sum_{j=1}^{256} c_j^{(\Delta MFCC)} N^{(\Delta MFCC)}(\mu_\mathbf{j}, \Sigma_\mathbf{j}; \mathbf{O_t})] \times [\sum_{k=1}^{128} c_k^{(E\&\Delta E)} N^{(E\&\Delta E)}(\mu_\mathbf{k}, \Sigma_\mathbf{k}; \mathbf{O_t})] \quad (1)$$

where s_m stands for the m^{th} state and $\mathbf{O_t}$ is a feature vector at time t.

The recognizer is a network of HMMs (λ_i) each representing a speech unit including silence. Each utterance must start and end with a silence. Transition between any two speech units is permitted, with a suitable transition probability implied which is zero if the transition is prohibitive by the phonetics of the language.

4. CHOICE OF SPEECH UNITS

Various choices of speech units as fundamental elements of a continuous speech recognizer have been investigated. They include phones, syllables, demisyllables, diphones, pseudo-diphones, and triphones. The general strategy is to seek more context specific representations in order to improve the resolution of the recognizer. The limitation to pursue such an objective is the inadequacy of training data for every unit. Take continuous Putonghua as an example, there are only 1210 left- and right-triphones, but there are 9026 triphones. Most of the triphones hardly appear in natural speech, thus imposing a difficult problem in building their training corpus and estimating their model parameters. A successful strategy is to adopt a new kind of speech units called generalized triphones. The idea is to group together phonetically similar triphones and represent them with a common hidden Markov model. Unseen triphones as well as those with inadequate training data will therefore be tied to triphones of ample training data. Thus, the former will get approximate representations, while the latter will have their model accuracies compromised inevitably. Sharing senones instead of sharing whole triphones mitigates but does not eliminates this problem. Furthermore, clustering triphones into a generalized one or clustering triphone states into a sharable senone requires the knowledge of an expert phonetician. In this paper, an inadequately trained triphone (with the number of training samples below a threshold) is approximated by a concatenation of its left- and right-triphones. Recognition accuracy as a function of such a threshold is displayed in Table 1.

5. A SEMI-TRIPHONE BASED CONTINUOUS PUTONGHUA RECOGNIZER

There are only 1210 left- and right-triphones in continuous Putonghua. It is therefore quite feasible to design a training corpus that includes all semi-triphones that can appear in a natural Putonghua utterance. Enforcing the appearance of a semi-triphone in an utterance by asking the speaker to read a specially designed semantically non-sensical syllable string can be done but the utterance is necessarily unnatural and should be avoided. As a consequence, there are about 100 unseen semi-triphones in each of the twenty speaker specific corpora. They appear mainly in transitions from a syllable to syllables /o/, /ei/ and /eng/ respectively. Their absence imposes no threat to the system performance because such syllabic combinations are the least expected in Putonghua anyway. Each semi-triphone is modeled by a left-to-right HMM of 2 states, preceded and succeeded by a non-emitting state respectively. Each state can transit to itself, the next state or the non-emitting state at the end of the model. Unseen semi-triphones have their state parameters determined by interpolation. Thus, a network of semi-triphones can be regarded as a triphone network with every triphone approximated by two concatenated semi-triphones. All semi-triphones are shared. Since a speaker may pause between syllables in the course of an utterance production, pauses change the speech unit sequence of the utterance and that has a serious effect on the training process. Instead of hand picking these unintentional pauses and informing the Baum-Welch training algorithm accordingly, a sequence of speech units compatible with the orthographic transcription of the utterance is supplied to the training algorithm. Whenever there is an inter-syllable transition among these speech units, an alternative path incorporating a /silence/ in parallel to the transition is constructed to cope with the possible existence of an unintentional pause as illustrated in Fig. 2.

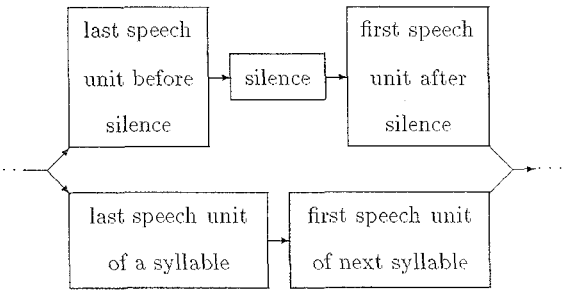

Fig. 2 - Alternative path to cope with an unintentional pause between syllables

6. HYBRID SYSTEM OF TRIPHONES AND SEMI-TRIPHONES

Since some triphones are much more popular than others in Putonghua, as in any natural language, they have much more training data than others. One cannot help wondering whether such popular triphones should not be explicitly represented by specific HMMs instead of being approximated by semi-triphone con-

catenations. A tied mixture HMM of 4-states is therefore constructed for each of those triphones with the number of training samples above a threshold. Such an HMM provides an alternative path in the HMMnet of speech units in parallel of the approximating semi-triphone concatenation.

Such a parallel architecture inevitably complicates the network of speech units and slows down the recognition process to some degree. However, it does improve the recognition accuracy of the system. A series of threshold values have been tried for several speaker dependent recognizers and the corresponding phone recognition accuracies of a typical system are tabulated in Table 1. The number of semi-triphones is maintained at 1210 in all cases.

Threshold	*No. of triphones*	*No. of states*	*Accuracy %*
∞	0	2419	82.56
260	11	2463	83.87
120	31	2543	84.67
70	63	2671	85.20
50	102	2827	85.59
30	206	3243	84.79
10	718	5291	85.59

Table 1 - Typical performance of a speaker dependent system with different number of triphones as a function of the threshold on the training sample size

The threshold value of 70 is selected as a compromise between recognizer accuracy and speed (proportional to the HMMnet size in terms of the number of states. That amounts to an average of 72 triphones in a sub-system. On the average, each speaker produces approximately one hour of speech data, 95% of which is used for system training and the remaining 5% used to test the accuracy of the system. The average accuracy is 88.2% with a standard deviation of 1.75%. Experiments have also been conducted to evaluate the merit of having the parallel concatenated path as an alternative in terms of recognition accuracy. It is noted that keeping both paths gives slightly better system accuracy than having the concatenated path of semi-triphones deleted.

Some Putonghua syllables are phonetic prefixes to other syllables, e.g. 'ji' is a prefix to 'ang' in 'jiang' and 'wu' is a prefix to 'en' in 'wen'. In order to segment a sequence of triphones into syllables unambiguously, the inter-syllabic and intra-syllabic versions of a triphone are treated as if they are distinct speech units. The notation for the inter-syllabic version of a speech unit bears a superscript * for identification. So, a triphone sequence like: *silence, j(silence,i), i(j,a(ng))*, a(i,ng), ng(a,silence), silence* should be decoded as syllables 'ji' and 'ang' bracketed by two silences. On the other hand, *silence, w(silence,u), u(w,e(n)), e(n)(u,n), n(e(n),silence), silence* should be decoded as 'wen' preceded and succeeded by a silence.

7. THE ENROLLMENT PROCESS

The enrollment process for a new user is to identify among the 10 speaker-dependent systems trained for speakers of the same gender as that of the new user, the one that performs with the highest recognition accuracy on a small set of speech samples from the new user. This small set of speech samples includes the 11 words of 2 to 4 syllables each and the 16 digit strings of 4 to 7 digits each as specified in the section about the speech database for system training above. Thereafter, the prototype is retrained with speech data from this new user as below:

Initially, this system cannot be expected to perform with high accuracy. The user will be asked to use the prototype and correct all mis-recognitions in the speech data inputed which are then saved for periodic system re-train. The prototype is re-trained by means of the Baum-Welch algorithm using the prototype parameter values as initial values. Choosing each of the 20 speakers as a new user in turn, the average recognition accuracy of the other 9 systems on the small set of speech samples mentioned above ranges from 58% to 75%. The test speaker has about 1 hour of speech data, half of which are used for testing while the other half are divided into 5 blocks of equal sizes for system re-train. Each block corresponds to approximately 6 minutes of speech data. Table 2 contains typical recognition accuracies of a re-trained prototype as a function of the amount of re-training.

Amount of retraining data in blocks	*Recognition accuracy %*
1	73.20
2	78.56
3	81.98
4	85.78
5	87.13

Table 2 - System performance as a function of retraining

8. POST PROCESSING LANGUAGE MODEL

In this recognizer, a word-class bigram statistics language model supported by a large vocabulary lexicon is used to convert a speech unit sequence into syllables and finally Chinese characters. The lexicon with 85,855 word entries covers virtually all the Chinese words currently in use. Each word entry is associated with 4 attributes: the GB code for the characters composing the word, the phonetic symbols for the syllables (Pin Yin), the word-class, and the prior probability of the word. In order to determine the word-class membership and prior probability of each word, as well as the bigram statistics between any two word-classes, a huge Chinese text corpus of over 63 million characters has been acquired. Because different from English texts, there is no explicit word marker in Chinese texts. In order to gather the above statistics, an effective Chinese

word segmentation algorithm by "Maximum Matching and Word Binding Force" [7], is applied to segment the text lines in the corpus into words.

The recognizer can generate the top n most likely speech unit sequences for each utterance to be recognized for the language model to post-process, but identifying the n best candidate sequences is time consuming. Most of the time, only a few of the speech units have multiple candidates unless n is very large. So, instead of asking the acoustic recognizer to produce n candidate sequences, only the top one is asked for.

For each phone p_i, close test results recognizing the training data produce a set of phones that have been identified to it, correctly or incorrectly. This set will hereinafter be referred to as the "confusion set" of p_i. In the following experiment, only the top 2 candidates in the confusion set are employed as possible candidates for p_i.

Hence, for each input utterance, only one phone sequence is required from the acoustic recognizer, and the confusion sets of the phones in the sequence can provide other possible phone candidates to the language model. The phone sequence is first converted into a syllable sequence. Phones from the beginning of the utterance or the end of a syllable are examined to see if they can form a syllable or the prefix of one. If a syllable can be formed, the conversion process carries on starting from the next phone. Whenever that fails, alternative phones from the corresponding confusion sets of those phones under examination will be considered.

The next step is to convert the syllable sequence into a word sequence. Because there are many homonyms in Chinese, the mapping of a sequence of syllables to a word is often one-to-many. For each syllable, a confusion set of 2 members is similarly constructed. If there are n syllables in the output sequence, a syllable matrix with $2n$ elements is built. Many syllable sequences can then be formed and each sequence may correspond to more than one character sequence due to homonyms and inexplicit word boundaries.

If a sequence of syllable images $\mathbf{o}_1, ..., \mathbf{o}_T$ is segmented into $\mathbf{o}_1^{w_1}, ..., \mathbf{o}_{k_1}^{w_1}, \mathbf{o}_1^{w_2}, ..., \mathbf{o}_{k_2}^{w_2}, ..., \mathbf{o}_1^{w_h}, ..., \mathbf{o}_{k_h}^{w_h}$ correpsonding to a word sequence of $w_1, ..., w_h$ which in turn, belonging to word classes $s_1, ..., s_h$ respectively, the soundness of the seguence is measured in terms of:

$$L = p(s_0 \mid s_h) \prod_{i=1}^{h} p(s_i \mid s_{i-1}) p(\mathbf{o}_1^{w_i}, ..., \mathbf{o}_{k_i}^{w_i} \mid s_i) \quad (2)$$

where s_0 is a word class of punctuation symbols appearing before and after the sequence of syllable images. $p(s_i \mid s_{i-1})$ and $p(s_0 \mid s_h)$ can be collected from the segmented text corpus while the probability of $\mathbf{o}_1^{w_i}, ..., \mathbf{o}_{k_i}^{w_i}$ forming a word w_i in s_i is defined as:

$$p(\mathbf{o}_1^{w_i}, ..., \mathbf{o}_{k_i}^{w_i} \mid s_i) = p(w_i \mid s_i) \prod_{j=1}^{k_i} p(\mathbf{o}_j^{w_i} \mid y_j^{w_i}) \quad (3)$$

Here, word w_i is a syllable sequence $y_1^{w_i}, ..., y_{k_i}^{w_i}$. $p(w_i \mid s_i)$ is the prior probability of word w_i, and $p(\mathbf{o}_j^{w_i} \mid y_j^{w_i})$ is the confusion probability to measure of the similarity between the observed image $\mathbf{o}_j$ and the syllable $y_j^{w_i}$ of the word w_i. Among all the possible word sequences, the principle of dynamic programming is employed to determine the optimal word sequence which has the maximum L.

With such a language model, the syllable accuracy is 86.8%. The character accuracies without and with recognizing the tones of each syllable are 70.2% and 77.1% respectively.

9. CONCLUSION

In this paper, a multi-speaker continuous Putonghua recognizer composing of 20 speaker-dependent recognizers as sub-systems has been presented. A large Putonghua database is constructed and used to train and test the recognizer. A language model based on word class bigram statistics is developed to convert speech units to Chinese characters. Confusion sets of phones and syllables are employed to improve the accuracy of the recognizer. The system covers a very large vocabulary and can be applied to various application areas.

REFERENCES

[1] L. Lee, C. Tseng, H. Gu, F. Liu, C. Chang, Y. Lin, Y. Lee, S.L. Tu, S.H. Hsieh and C. Chen, "Golden Mandarin (I) - A real-time Mandarin speech dictation machine for Chinese language with very large vocabulary", IEEE Trans. Speech and Audio Processing, Vol. 1, No. 2, pp. 158-179, 1993.

[2] L.S. Lee et al, "Golden Mandarin (II) - an improved single chip real-time Mandarin dictation machine for Chinese language with very large vocabulary", Proc. of IEEE ICASSP, pp. 503-506, Minneapolis, 1993.

[3] H.W. Hon, B.S. Yuan, Y.L. Chow, S. Narayan and K.F. Lee, "Towards large vocabulary Mandarin Chinese speech recognition", Proc. of IEEE ICASSP, pp I545-I548, Adelaide, 1994.

[4] Y.Q. Gao, H.W. Hon, Z.W. Lin, G.Loudon, S. Yogananthan and B.S. Yuan, "Tangerine: a large vocabulary mandarin dictation system", Proc. of IEEE ICASSP'95, pp. 77-80, Detroit, 1995.

[5] L.S. Lee et al, "Golden Mandarin (III) - a user-adaptive prosodic-segment-based Mandarin dictation machine for Chinese language with very large vocabulary", Proc. IEEE ICASSP, pp. 57-60, Detroit, 1995.

[6] L.S. Lee et al, "Complete recognition of continuous Mandarin speech for Chinese language with very large vocabulary but limited training data", Proc. IEEE ICASSP, pp. 61-64, Detroit, 1995.

[7] Pak-kwong Wong and Chorkin Chan, "Chinese Word Segmentation Based on Maximum Matching and Word Binding Force", Proc. of 1996 International Conference on Computational Linguistics, pp 200-203, Copenhagen, Denmark, August 5-9, 1996.