Speaker-independent Dictation of Chinese Speech with 32K Vocabulary

Bo Xu Bing Ma Shuwu Zhang Fei Qu and Taiyi Huang

Speech Research Group, National Lab of Pattern Recognition
Institute of Automation,Chinese Academy of Sciences
Beijing 100080 , P.R.C
Email : xubo@prldec3.ia.ac.cn

Abstract

While early machines adopted isolated syllable as input units and needed boring enrollment, our research focus on the speaker-independent, word-based dictation. A deliberately designed 120-speaker database was built for training ; inter-syllable context ,tonal and endpoint dependent acoustic model are applied with promising MFCC feature; Two-pass acoustic matching accelerates the recognition making fully advantage of the monosyllabic structure of Chinese speech; A complete word bigram and trigram serve as language processing module. With all efforts, the system reaches 90% character accuracy performing in almost real-time on Pentium PC without DSP help.

1. Introduction

For non-alphabetic languages like Chinese, official dictation emerges as most potential application for speech recognition. Benefited from the monosyllable structure of Chinese language, early dictation machines adopt isolated syllables as input units[1,2]. Though the way can extend to the unlimited vocabulary recognition only inserting new words to lexicon without revising any component of acoustic decoder, it is obviously unnatural for user to speak syllable by syllable; Since 1992, reports on the polysyllabic words recognition are appeared[3,4,5]. The new proposed utterance way alleviates the above shortcoming greatly,but they still need to speak several thousand words as a procedure of enrollment. In this paper, we will focus on some aspects of the speaker-independent, word-based dictation,which is taken as one of middle step towards speaker-independent all speed dictation technically. A 120-speaker database(each uttered about deliberately selected 160 syllables and 300 words) for acoustic training was described ; In building the HMM, Some advanced features such as the inter-syllable coarticulation, tone effect and position of syllables in the words are considered together with promising MFCC representation of features; Two-pass acoustic matching accelerates the recognition making fully advantage of the monosyllabic structure of Chinese speech; A complete word instead of the word class bigram and trigram serve as language processing module. With all efforts, the system reaches 90% character accuracy performing in almost real-time on Pentium PC without DSP help. In the following sections, each topic will be discussed independently.

2. Multi-speaker Database “COSDIC”

To make the database widely accepted and extensively used,we deliberately designed the training script. The database mainly consists of three parts: monosyllables, connected digits and words or phrases. Three criteria are setup for the selection of the specific syllables and words from 1240 tonal syllables and 45K commonly used words which is totally impractical to be all read for every person. Firstly we consider enough templates of coverage on possible connections of two modeling unit. Currently two kinds of modeling units is considered,namely Initial/ Final and phonemes. The former one is adopted mostly in state-of-art Chinese speech recognition system and the later one is being explored in some researches(such as in our group and HongKong University). Secondly, we choose the speakers from the various region of China to ensure the coverage of main dialects. Thirdly, connected digits are recorded for every person considering all digits connection. In fact, every person’s utterance has enough templates of single digits and connections of two digits,It is relatively a independent database which can be used for connected digits recognition purpose. Finally, top 100 most commonly used function words is sampled as a special part of the database. The scale of database reaches 120 peoples(60 males and 60 females) with every about 160 syllables, 300 words and 70 digits string. The speech is recorded under 38-42db SNR background through the general sound card for multimedia purpose which is consistent with the input channel of online system. When initial and final parts are taken as modeling units,we have 30 templates for every initial and final connections.When phonemes are taken as

modeling units, 120 templates are supplied for every diphone training.

3. Language Model Building

The task we attack is the recognition of the largest newspaper in China "People's Daily". we collected 18 millions characters of original texts for setting up two levels of language model. The first level language model is the syllable bigram disregarding the tones of character .The syllable-level language model is based on the labeling of the pronunciation to every characters in the corpus. The distributions of syllable frequency (syllable unigram) and syllable connection(syllable bigram) of every syllable are rather appealing which are expected to reduce the perplexity of syllable-level grammar. Because there is no space between words , the first step to build the second level word bigram language model is to make the segmentation of words in the original texts. We compiled an initial dictionary with 65,000 words and process about 1/10 corpus man-machine interactively to add new words into the dictionary and insert a space between every word. After the formation of large enough dictionary , the left 9/10 corpus is processed total automatically. Obtaining the segmented corpus , we then built word bigram (actually it is long distance character n-gram, n is more than 4) . We select the most frequent used 320,00 words(which are occurred more than 5 times in corpus)as basic dictionary. Other words may be combined with the single syllable words.We use those two level language models for later multi-pass searching.Since there exist some ambiguities in the word definition and segmentation in Chinese language, we process the corpus iterately to minimum the error.

Unlike western languages, there is no concise definition of what a word is and this ambiguity of words definition causes the end-user rather difficult to speak properly and skillfully for realistic application. This problem only can be thoroughly solved in continuous speech recognition and any other out-of-vocabulary words can be combined with the words in the basic dictionary. Those are the benefits of continuous speech. However in this research, some methods are adopted to alleviate this problem. As we know, most people are rather puzzled by some fixed connections of words,whether they are a word or two words. So in our processing, after one pass of segmentation, two monosyllabic words with very large connection probability are collected and defined as a new bisyllabic words, we also adjust segmentation based on the calculated bigram and trigram.

4. Inter-syllable Context ,Tonal and Endpoint Dependent Acoustic Modeling

It is well known that Chinese is a monosyllabic structured language. Some research interest has already been focused on modeling the intra-syllable coarticulation effects and has been shown to be very effective.However the study of modeling co-articulation of inter-syllable was usually ignored;Chinese is also a tonal language. For Putonghua,there are basically four lexical and a neural tone. Each syllable can be specified by a certain combination of an initial and final assigned with one of five basic tones. Conventionally, finals are modeled disregarding its tones; For a word recognition system, high frequency used words is monosyllable words and it is always is ‘contaminated’ by the inter-syllable coarticulation if it is mixed with polysyllabic words. We also try to have a small variance of models at the first layer of lexicon tree. Based on those reasons, we try the inter-syllable context, tonal dependent and endpoint dependent acoustic modeling.

4.1. Inter-syllable context dependent

In the case of continuous speech recognition of Chinese,all the initial and final parts are affected by both left and right phonetic components. Considering the influence from right context,certain right context final models were added to 100 intra-syllable initial models. In inter-syllable models the phone groups were used instead of phones. The initials of the next syllable can be divided into 7 groups according to the manner of articulation of initials:

Stops: /b/,/p/,/d/,/t/,/g/ and /k/
Fricatives: /f/,/h/,/x/,/s/,and /sh/
Affricates: /z/ ,/zh/, /c/, /ch/, /j/, and /q/
Liquid: /l/
Nasal consonants: /n/
Nasal consonants: /m/
Voiced consonants: /r/

If the next syllable is a null initial it can be divided into 6 groups based on their first vowel. Now we have 100 initial models and 38*13 right context-dependent final models.

Considering the influence from left context only, 21*10 initial models and 38*8 final models are needed. If the left context is finals we divided the left context of into 10 groups based on the last phoneme in the final of the preceding syllables. If the left context is an initial, we divided the left context into 8 groups(7 groups are real consonant initials and one group is silence-initial). The following show the error rate reduction with the different acoustic models.

Table1: Inter-syllable context acoustic models

Type of Models	Number of Models	Error Reduction
context-independent	22+38	----------
left context-dependent	21*10+38*8	20.3%
right context-dependent	100+38*13	52.4%

From the table we can conclude that the initial/final parts are influenced by both left and right context, but the influence from right context is much larger than that of left context. Therefore the right context-dependent models are more effective than the left context-dependent one; The intra-syllable is just a special case of right context-dependent models only considering the initial parts. So we are sure that inter-syllable context-dependent models are necessary for word and continuous speech recognition.

.4.2. Tonal dependent model

Since Chinese is a tonal language, the tonal cues play a very important role in word recognition. It affects the pitch tendency ,energy and duration of final part. A test has been conducted with the recognition system same as the above. In the test, 100 intra syllable context-dependent initial models and 38*5 tone-dependent final models are built. The result show that tonal information is very important in building acoustic models, in fact it can be thought as extra features on the MFCC dimensions.

Table2 : Tonal-dependent acoustic models

Type of Models	Number of Models	Error Reduction
Intra-syllable context dependent(only initial)	100+38	----------
tonal dependent (final)	100+38*5	31.1%

4.3. Endpoint dependent model

As we noted ,when monosyllables are recognized under the task of words recognition(1-7 syllables),the recognition accuracy degraded greatly. It can be explained that the “pure” initials and finals in monosyllables is contaminated by the initials and finals in the polysyllabic words. There are two reasons to modeling those “pure” separately. Firstly, a lot of most frequent used words are single syllable words and the correct recognition of those words will affect the performance of system greatly. Secondly, As we construct the lexicon in tree, the parsing of the first layer nodes of the that tree is crucial for fast and accurate recognition. So we proposed endpoint-dependent acoustic modeling to distinguish those models at the beginning or tail of the words. The experiment show that this modeling methods improve the recognition greatly.

Table3: Endpoint dependent acoustic models

Type of Models	Number of Models	Error Reduction
intra-syllable context dependent models	100+38	--------
endpoint dependent models	100*2+38*2	34.2%

5. Multi-pass Search

There are many strategies to apply the multi-pass search. In English recognition system ,most of them pass the lexicon tree constrained with word bigram as the first-pass search. However, Chinese is monosyllable-structured language and there are only 407 pronunciation if disregarding its tones. This attributes make it possible to construct a relatively simple acoustic decoder which is independent of the vocabulary and utterance mode. This is also the advantage of Madarian over west language in automatic speech recognition. We test a lot of strategies and the following proposed three-pass recognition algorithm is the most efficient one.

5.1. First pass on syllable tree

The first pass is running in a static tree-structured syllable concatenation. Since the aim of the pass is to provide acoustic heuristic information , the simple intra-syllable acoustic model with no grammar constraint is embedded. This allows for the structuring the network in a tree and significantly reduces the computation. This pass outputs a list of initial and final parts and associated scores possibly active at the certain time of the input speech. At this pass, we record the set of active HMM reaching the final state at t of HMM as $W_t=\{H_{t1},H_{t2},....,H_{tn}\}$, Its corresponding probability as $P_t=\{P_{t1},P_{t2},....,P_{tn}\}$,we use set Ω_t as the possible start point at time T-t when applying the second backward pass.

5.2 Second pass on Lexicon Tree

The second pass is running very quickly on the reversed lexicon tree which represents 32K vocabulary embedded with inter-syllable or endpoint dependent acoustic model. During the the Viterbi-beam search, We only expand those models in Ω_t at time T-t. Also we execute pruning according to the heuristic probability when grammar nodes jumping occur. This two-pass acoustic matching take fully advantages of the monosyllable-structure of Chinese language while keep the worth of the lexicon constraint.

5.3. Third pass on Language Model

After acoustic processing,we have a word list for every utterance. Here we have most 10 candidates and every candidates may have up to 50 homonyms. puntuations are taken as one vocabulary in the lexicon. In our isolated recognition system, the language model is applied as a post-processing process to the acoustic analysis. Here another pass of Viterbi beam search is applied with the every step forward of word utterance in the whole sentence. The combined scores are obtained from the acoustic scores and all included language model scores:

$P_{total}=C1Pa+C2Pu+C3Pb+C4Pt$

Herea Ci, i=1,2,3,4 are the weights for acoustic,unigram,bigram and trigram scores.

6. System Experiment

The speech is sampled at 16khz and 12 order MFCC(Mel-scale frequency cepstral) ,delta MFCC,two-order delta MFCC and log energy are extracted. We compare LPCC and MFCC in various tasks (such as all syllables recognition, speaker independent recognition, noise resistant experiment etc.) . All results show the superior characteristic of MFCC than LPCC, averagely 10% -- 40% error reductions which dependents on the baseline recognition accuracy. Transformation pursuing the input channel independent features are operated. Then Discrete HMMs are used as modeling tools with 3 distributions for every units. The system are transplanted to Windows95 and can receive speech inputs continuously while displaying the updates recognition results on the screen. Now more advanced technology are being integrated into the system.

we test 120 sentences which were provided by the third party as the formal evaluation organized by National High Tech Research & Development Project. Each sentences consists of 10 to 40 words which are recorded through unknown input channel(including microphone and AD). Following are the summary results of test.

Table.4: Character and Word Accuracy

	Word Accuracy	**Character Accuracy**
Top-1	**65%**	**90%**
Top-10	**96.2%**	**---**

We are pleasant to find that when we have 96% top-10 inclusion rate of acoustic matching, the accuracy of character conversion can reach to 90%, which is general only 4-6% differences. The result demonstrate the complete word statistical model(especially trigram) is much powerful and precise than word-class based one.

7. Main References

1. Gao, et al, "A real-time Chinese speech recognition with unlimited vocabulary", ICASSP pp.257-260

2. L.S.Lee et al, "A Real-time Mandarin Dictation Machine for Chinese Language with Unlimited Texts and Very Large Vocabulary", ICASSP'90, pp65-68

3. Eng-Fong et al, "The use of tree-trellis search for large-vocabulary Mandarin polysyllabic word speech recognition", Computer Speech and Language(1994)

4. Hsiao-Wuen Hon et al, "Towards large vocabulary mandarin Chinese speech recognition", ICASSP-94 I-545-I546

5. B.Xu , et al, "A 46500-word Chinese speech recognition system",ICSLP'92

6.Steve Austin,et al, "The Forward-Backward search algorithm", ICASSP'91 p697-700

7.Frank K. Soong et al, "A Tree-Trellis Based Fast Search for Finding the N-best Sentence Hypotheses in Continuous Speech Recognition", ICASSP' 91 p705-p708