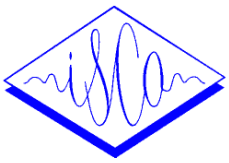

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

A 46,500-word Chinese Speech Recognition System

B.Xu,Z.W.Lin,T.Y.Huang,D.X.Xu & Y.Q.Gao

National Lab of Pattern Recognition
Institute of Automation
Chinese Academy of Sciences,Beijing 100080

Abstract: This paper introduces a large vocabulary isolated words speech recognition system for Chinese speech based on HMM modelling of phonemes.Through the concatenation of phonemes a word HMM model can be formed in training stage,in this case the coarticulation of connected utterance is preserved in the model.In recognition phase,Syllable String Network(SSN) without any lexical constraint outputs one or two phonetic strings with different length,then every word in library is matched and scored comparing with these one or two strings.Taking 50--100 candidates,rebuilding lexical tree,a less-pruning Viterbi Beam Search(VBS) is applied to get final result.The system achieves 86% recognition rate for top-1 and 94% for top-5.A concept of using pre-computed confusion matrices is proposed for phoneme string mathcing in this paper.Also the way of estimation of these matrices is provided..

10.21437/ICSLP.1992-205

1.Introduction:

Chinese language has some distinguishing features,such as,every Chinese word consists of one or several Chinese characters,the pronounciation of each character is a monosyllable,and the total 1300 monosyllables cover the pronounciation of all the Chinese characters.Therefore ,many researchers have been interested in the scheme of the recognition of all these monosyllables to reach unlimited vocabulary recognition and a lot of real-time systems have been set up as dictation machines.But ,there are still two barriers preventing them from real application.One is its poor naturality for the speaker to say a word syllable by syllable.The other,especially of vital importance for language processing,is its hardness of automatic word segmentation in Chinese.Utterance in word mode can overcomes above two shortcomings.

However ,a lot of recognition difficulties emerge when syllables are connectedly spoken.For example,the boundary between syllables are uneasy to locate,severe coarticulation lead to large number of phoneme variabilities and memory storage and computation increase tremendiously.Some unusual way which is different from syllable recognition must be explored to complete this task.

According to the statistics,there are about 46,500 most commonly used words,including 5070 1-syllable words,31187 2-syllables words,5125 3-syllable words,4566 4-syllables words 382 5-syllables words,144 6-syllables words and 46 7-syllables ones.In following discussion,we will concentrate on the recognition of these words and outline the main aspects of the first system of this kind to realize the dream of inputting texts in speech.

2.HMM modelling

2.1 Speech Processing

The speech is sampled at 10khz with 14 bits accuracy and pre-emphasized with a filter of 1-0.97z-1.Then,a Hamming window with a width of 25.6 msec is applied every 12.8 msec.Auto correlation analysis with order 12 is followed by LPC analysis with order 12.Finally,12 LPC-derived cepstral coefficients and differential coefficients are calculated and quantitized individually with two sets of codebook.The codebook size is 128.

Endpoint and silence detection are obtained from energy and zero-crossing information.During sampling period there is a circular memory data structure saving all frame parameters necessary for next processing,,so endpoints can be moved forward or backward freely to fit the various noise background.

2.2 Word HMM model

In Chinese,every syllable has regular CV or V structure,here C is refered as initial part and V is refered as final part.All initial part are consonants All final part are vowels or vowel plus nasals.There are 22 initials and 37 finals totally written as:

22 initial parts:{ 0,b,d,g,p,t,k,l,m,n,r,j,z,c,s,zh,ch,sh,q,x,h,f}
37 final parts: { a,an,ai,ao,ang,e,ei,en,er,eng,
o,ou,ong, ü,üe,ün,üan,
u,ua,ui,un,uo,uan,uai,uang,
i0,i1,i2,ia,ie,in,iu,ian,iao,ing,iang,iong }

here symble "0" represents the V structure situation.Every word consists of several characters and may be written as C1V1C2V2...CnVn.In real word's HMM,we insert two kinds of models to represent the silence part between syllables.One is called "sil",inserting before syllables begin with plosive{b,d,g,p,t,k,h},another is called "ops",inserting before syllables begin with the other initial {j,z,c,s,zh,ch,sh,q,x,f},which may be optional in model and a transition with probability 0.5 can jump over it .Since there is no silence signal part before voiced consonant and vowel,so there is no any "sil" or "ops" insertion before {l,m,n,r} and vowel (initial is "0").

For example,HMM model of word "zhong guo ke xue yuan",means Chinese Academy of Sciences, is the concatenations of 12 phoneme HMMs.

"zh"+"ong"+"sil"+"g"+"uo"+"sil"+"k"+"e"+"ops"+"x"+"üe"+"üan"

2.3 The choice of training set

When syllables are connected spoken,latter utteance may have influence on the previous one acoustically,the influences between intrasyllable consonant and vowel also existed.This language phenomenon must be reflected as throughly as possible during training stage.So training set choice is very

crucial to the performance of the system.We mainly keep following three rules in automatic generation of this set:

.Adequate coverage of all syllables(also initial and final parts)

.Adequate coverage of all vowel to all consonant juncture

.And especially important,adequate coverage of junctures of vowel with those voiced consonant {m,n,l,r} and vowel(the syllable with initial "0") itself .Our experiences show those words with no silence signal boundary between syllables are most difficult to recognize.Hence,in trainig stage those junctures must be emphasized.

2.4 The initialization of phone HMM

For discerete density pdf's,the complex initialization algorithms are usually not necessary,but good initial model may speed training procedure greatly.Since the availability of monosyllable utterance,we chose 300 syllables as initialization material.The short time energy and zero-crossing rate for every frame is used to determine the voiced/unvoiced segmenting point ,then all phoneme models are initialized by its corresponding speech signal segments.Result shows this bootstrapping method is very efficient and only one iteration of Forward-Backward procedure is enough for the complete model training in our system.

3.Three Stage Recognition

It's time-consuming to do Viterbi-beam search(VBS) directly on the 46,500 lexical tree.But any attemps to prune some states using only local acoustic information can lead to irrecoverable errors.Noticing the substituted errors output by all recognition system,which are rather confusive in acoustic features with target ones ,we proposed a three stage recognition method named "Phoneme String Searching in SSN","String Maching" and "Detailed Matching" ,respectively.

3.1 Phoneme String Searching in Syllable String Network(SSN)

There are many ways to construct grammar network for Viterbi-beam-searching(VBS).The most complete and complex one is the lexical tree of total 46,500 words with which every path from root node to leaf node is a valid string of word.Though it can achieve very high recognition rate,it is too time and memory consuming and never be used for large vocabulary recognition.Another option is to construct syllable bigram network(concerning partial lexical information) basing on the statistics of whole vocabulary.The total grammer node in this case is 5600(including node of "sil"or "ops") with 62000 links,however ,this network always generates many invalid string , e.g. [a b c] is valid path ,[e b f] is also a valid path,the network may output invalid path [a b f] or [e b c].These strings resemble correct word very much in acoustic characters and a postprocessing can easily recover them to proper one .Unfortunately,it still cost too much in computation and almost impossible to be implemented under PC enviorment.This situation leads us to try a more simple network called Syllable String Network(SSN) without any lexical constraint.In our system,SSN is just concatenated with 7 syllable nodes along with 7 ending nodes,each syllable node contains only 400 links between 22 consonant node and 37 vowel's.The number 400 is just the number of total non-tonal monosyllabes in Chinese.The main advantage of this network is that least pruning is feasible in SSN searching. This avoid irrecoverrable errors due to the unmatured pruning using only local acoustic information for speed reason.The SSN is set multi ending nodes for two reasons.First ,to represent words with different number of syllables in library,secondly,to get two phonetic strings with different length after fisrt-stage seaching,avoiding the time-cost dynamic string mathcing.

In this stage's VBS,we keep about 80 active states per frame.SSN has 60 grammer nodes per syllable node,every HMM has 4 states,so one syllabe node has total 240 HMM states.80 active states per frame are enough to avoid inproper pruning while keeping enough quick search.

3.2 String Matching for Preselection

The string matching is the second stage of the whole recognition and the key step for fast lexical information access.The task of this stage is to compute similarities between a phonetic string and each dictionary word,represented in terms of phonetic symbols.Different from the other system,the similatity is defined only through the cost of substitution,excluding any deletions and insertions cases.We calculate scores of each words in library with the same length of the output phonetic string individually.If the SSN's output of one CV string is C1V1C2V2...CnVn,a word with n syllables in dictionary is C1'V1'C2'V2'...Cn'Vn',then we simply score this word as:

$$\text{Score}=\Sigma_{i=1}^{n}[M_1(C_i,C_i')+M_2(V_i,V_i')]\,/\,n$$

If another output of SSN is C1V1C2V2...CmVm,a word with m(m not equal n)syllables in dictionary is C1'V1'C2'V2'...Cm'Vm',similarily we score this word as

$$\text{Score}=\Sigma_{i=1}^{m}[M_1(C_i,C_i')+M_2(V_i,V_i')]\,/\,m$$

In formulas,M1 and M2 are pre-computed matrices.M1(22x22) represents the similarities between all initial parts,M2(37x37) represents the similarities between all final parts.The scores for words represent the similarities between every word and output assumed strings.Order the scores roughly and we get a short list candidates(from 50--100).Noticing that every score's calculation only cost 2n(or 2m) addition and one division,this method reduces the computation greatly by a factor of 20-30 and the computation complexity is almost independent of the vocabulary size.

3.3 Detailed Matching

Reconstructing the short list candidates to a lexical tree(every node is a phoneme HMM) ,applying the VBS again,we obtain the final 5 candidates.In most cases,the preselected words have a large common phonemes,so the lexical tree expands modestly ,thus gives us a chance to do less pruning viterbi search . Unlike the general grammar network,the path from root node to paticular ending node is unique,so some backtracing information can be ignored resulting in the less run time and less storage requirements.

A very practical strategy called Multi-Threshold Pruning is used througth the VBS searching.We know that keeping the proper number of active states is very difficult for a fixed threshold.Sometimes searching is so fast that very crucial states are lost while sometimes too little pruning makes the system looked as deadlocked.Keeping fixed active states

need ordering of all states' probabilities.The Multi-Threshold Pruning is the compromise solution to these problems.In our system,we set 8 threshold(T1....T8) and 9 probability block(negative huge,T1),(T1..T2)....(T7,T8),(T8,positive huge).First we check which block every state's probability fall into and count the number of states every block (Ti,Ti+1) have,we then can activated those states whose probabilities are larger than the threshold of one Ti in next frames searching and keep the proper number of active states every time.

4.The Estimation of Confusion Matrix

In section 3.2 we proposed a method to use confusion matrix for string match.Usually "confusion matrix" provides a way of measuring the quality of the phonetic recognizer.If diagnal number is greatly larger than nondiagnal's,we say this system performs well.On the other hand,this measuing also shows its similarities of these phones under the specific recognizer.As a result,it can be used for roughly matching betwen a phonetic string and each dictionary word. Comparing three sets of confusion matrices estimated from three persons individually,we find their data are rather close.Thus confusion matrices are possible to be speaker-independent.

Estimation of phonetic confusion matrices M1 and M2 needs a lot of confusive phonetic pairs actually generated by recognition system.For avoiding the comlexity of dynamic matching,we only keep one ending node for SSN to be sure only phonetic string with specific length is possibly generated.For example,when we feed 4-syllable words to SSN,only one ending node exists at the end of 4th syllable.This makes estimation very convenient. If SSN's output string is C1'V1'C2'V2'..Cn'Vn',feeded word string is C1V1C2V2...CnVn,then the pairs are counted as

count of confusion pairs(Ci,Ci')++; i=1.....n
count of confusion pairs (Vi,Vi')++; i=1,.....n
count of phoneme Ci++; i=1.....n
count of phoneme Vi++; i=1.....n

After all training data are processed,the confusion degree of phones P(C or V) and P'(C' or V') is estimated as:

M(P,P')= Total number of confusion pairs(P,P')/Total number of P

For further such pairs under limited available training data,we employed lattice N-best to get more candidates take part in estimation.For these more candidates,we use the time-synchronous forward-pass search algorithm,with only theory at each state,we add the probabilities of all paths that come to each state.At each grammar node (for each frame),instead of remembering only the best scoring word ,we store all of the different words that arrive at that node along with their respective scores in traceback list.At the end of the words,we simple search(recursively)through the saved traceback lists for all of the complete sentence hypotheses that are above some threshold below the best path.An example of 5-best may result:

zhong guo ke xue yuan --------- zhong uo he que yuan
zhong guo ke xue yuan --------- zong guo he xue yuan
zhong guo ke xue yuan --------- zhong guo ke que juan
zhong guo ke xue yuan --------- zhong duo pe xue yuan
zhong guo ke xue yuan --------- zhong duo he que yuan

Experiment shows larger number of confusion pairs make estimation more accurate and suitable to real recognition and also improve the robustness of the system.

To make the confusion degree more accurate ,words with different length may have its own matrices with a little increasing of memory storage and no increasing of computation.Also enough training data must be ready for such pairs generated.

5.System Preformance

The system is built on a PC-486 with a TMS320C30 DSP board.The TMS320c30 realizes the endpoint detecting and pre-processing of speech signal on line,main computation such as Viterbi-beam search and fast scoring are done by PC-486.The coverage of correct recognized word reachs 98% when size of candidate set in preselection is 100.After detailed matching,the top-1 accuracy is 87% while top-5 above 94% .For a 10072-word active vocabulary,the correct recognition rate is 91.4%.The system can response within 3 seconds .

Reference:

1.F.Jelinek,et al, "Experiments with the Tangora 20,000 Word Speech Recognizer", Proc.ICCASP 1987.Vol 1
2.V.N.Gupta,et al,"Using Phoneme Duration & Energy Contour to Improve Large Vocabulary Isolated-word Recognition", Proc ICASSP 1991,P 341- P 344
3.R.Billi,et," A PC-based Very Large Vocabulary Isolated Word Speech Recognition System",Speech Communication 9 (1990)
4.Richard Schwartz, " A Comparison of Several Approximate Algorithms for Finding Multiple(N-best) Sentence Hypothese", IEEE ICASSP 1991, P701--P704
5.Richard Schwartz, "The N-best Algorithm:An efficient and Exact Procedure for Finding the N-best Likely Sentence Hypothisize",IEEE ICASSP 1990 p81-- p84.
6.Chin-hui Lee,et al, "A Frame-Synchronous Network Search Algoirthm for Connected Word Recognition",IEEE Transactions on ASSP,Vol 37,1989