Use of Prosodic Information to Integrate Acoustic and Linguistic Knowledge in Continuous Mandarin Speech Recognition with Very Large Vocabulary

Hung-yun Hsieh, Ren-yuan Lyu and Lin-shan Lee

Dept. of Electrical Engineering, National Taiwan University
Taipei, Taiwan, Republic of China

ABSTRACT

This paper presents a new approach to use prosodic information for the integration of acoustic and linguistic knowledge in continuous Mandarin speech with very large vocabulary. Since the overhead computation incurred from unification of search space is confined to the syllable boundaries, the use of prosodic information to reduce the syllable boundary hypotheses as well as the syllable matching length is shown to be effective. The inherent complexity with the very large vocabulary is also reduced by the use of phrase boundary hypotheses conjectured via the phrase-final lengthening. Experimental results show a 47.2% recognition time save with only 5.67% error rate increase using the syllable and phrase boundary hypotheses conjectured from prosodic information.

1. INTRODUCTION

The recognition of continuous Mandarin speech with very large vocabulary has had a paradigm of separation of knowledge usage. Consider, for example [1][2], the speech recognizer with the two-staged architecture. First the syllable lattice is prepared by the use of Viterbi algorithm, then out of the syllable lattice a word lattice is constructed with the help of lexical access, and finally a linguistic decoder is employed to find the best word sequence as the recognition result. For another example [3], a coarse acoustic model is first applied to generate a larger syllable lattice, then after the word lattice is again constructed a detailed acoustic model is used together with the language model to search for the best word sequence. The success of these architectures in continuous Mandarin speech for fast recognition is obvious due to the plurality of information carried by the syllables, i.e. all the characters in Chinese are monosyllabic and the total number of syllables is quite limited. However, the potential loss herein is not negligible. Not only we lose the frame synchronism due to repeated optimization stages, but end up with local optima for lack of interaction among knowledge sources. The deficiency would be significant when the acoustic model is not reliable, and the propagation of errors floods as the number of stages increases.

In this paper, instead of constructing separate search spaces bridged by lexical access, the search with acoustic and linguistic models is unified in a common space spanned by the lexicon for one-staged recognition. Though more than 80,000 commonly used words are compiled in the lexicon, the complexity increase is moderated by the use of syllable and phrase boundary hypotheses conjectured from prosodic information.

2. INTEGRATION OF ACOUSTIC AND LINGUISTIC KNOWLEDGE

It is difficult to use same units both for acoustic and linguistic models in continuous speech with very large vocabulary, so the lexicon plays a critical role if acoustic and linguistic knowledge is to be unified in a single stage with efficiency.

2.1. Lexicon Organization

After transcribing all the lexical entries into their character sequences, the lexicon is ready to be organized as a forward tree data structure. The root is a null character and every other node of the tree stands for a character with a unique pronunciation label as **Figure 1** shows. Some nodes are terminals (circles with solid line) and some are not (circles with dashed line); any path starting from the root to a terminal stands for a lexical word. For the lexicon with

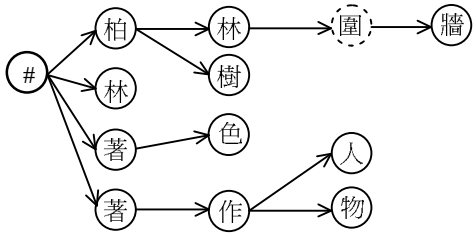

Figure 1: Tree lexicon structure

84,480 entries, the total number of nodes in the tree is 103,109, which means in average, for the 14,211 characters in the lexicon, every character is shared by 7.26 nodes due to different context inside the words. Compared with the search space constructed by 1,345 syllables in the two-staged architecture, the search space is now 76.66 times of it.

2.2. Search Strategy

The search for optimal word sequence is divided into two parallel modules: the lattice for syllable recognition, and the tree for word score accumulation. For the lattice module, a concatenated syllable matching (CSM) algorithm [2] is used such that recognition for isolated syllables is required for every possible syllable interval determined by LPC gain dips and syllable duration constraints.

This implies the search for the optimal path inside a syllable is not complicated in this search algorithm. For the tree module, the score of each word is accumulated node-by-node with the score of each node taken from the lattice module according to the tree structure. **Figure 2** shows such a configuration where the dashed lines

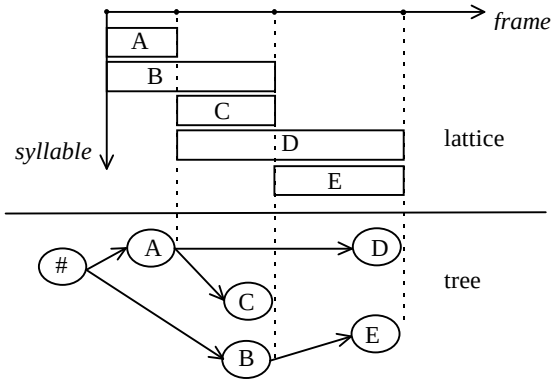

Figure 2: Search with two modules

indicate syllable boundary hypotheses. For all words reaching their terminal nodes, acoustic scores as well as the linguistic scores are transferred to the receiving words during the word transition. Unlike most other tree search algorithms [4], the incorporation of language model parameters is not delayed until the identity of the receiving words is known. By using word class bigram based on starting/ending characters [5], it is possible to apply exact word class bigram value at the word boundary even when the *exact* word that followed is not known, because words with same starting characters are grouped in a class. It is also noteworthy that since the score of the word is only important when being transferred to following words, the accumulation for word scores is thus delayed to the hypothesized syllable boundaries, instead of frame-by-frame computation. In this way, the burden of maintaining the word scores of the very large vocabulary size is amortized over the frame separation of adjacent syllable boundary hypotheses, which is to be minimized by the use of prosodic information.

3. USE OF SYLLABLE BOUNDARY

Though quantity of the syllable boundary hypotheses conjectured via LPC gain dips is satisfactory [2], the quality is not. Thus syllable boundary conjectured from other reliable sources will be used to strip off the unpromising dips for the purpose of speedup.

3.1. The Probability of Voicing

The particular INITIAL/FINAL structure of the Mandarin syllable is quite helpful as a cue of syllable boundary; because all the FINAL's in Mandarin are voiced speech, and most of the INITIAL's are unvoiced, the presence of unvoiced segment in the speech is presumed to be the syllable starting INITIAL's as well as the pause between syllables for breathing. Consequently a voiced/unvoiced decision, or a measure of the probability of voicing, is necessary if better boundary hypothesis is desired.

A real-time recurrent network (RTRN) [6] as **Figure 3** shows is

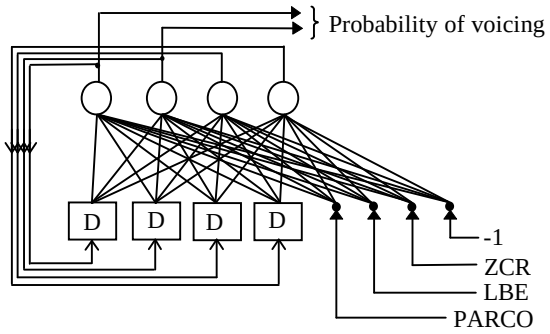

Figure 3: Real-time recurrent network configuration

adopted for the purpose of providing reliable voiced/unvoiced decision. Three features relevant to the probability of voicing are used as network input: zero-crossing rate (ZCR), low-band energy below 400Hz (LBE), and the first reflection coefficient from LPC analysis (PARCO). Output from a sophisticated pitch tracking algorithm [7] is taken to guide the network in the learning phase. Sample output from the well-trained network may look like **Figure**

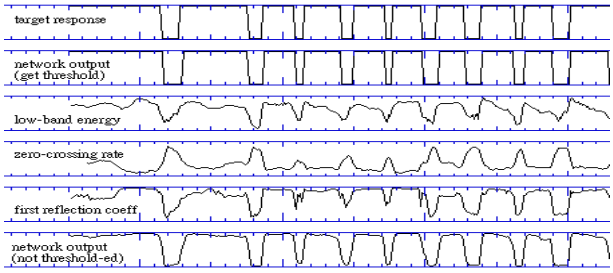

Figure 4: Sample output from the real-time recurrent network

4 where it is evident that despite the rugged contours of the three input features, the network output is quite smooth. Moreover, although the presence of some syllable boundaries may not be so prominent viewing from individual input feature, the information in collaboration gives a solid decision.

3.2. Pruning Techniques

After imposing proper threshold on the network output, the syllable boundary is set to the frame when contour of the probability of voicing falls from 1 to 0. Since boundaries thus obtained are presumed reliable, dips around these syllable boundaries within the reach of minimum syllable duration are no longer valid hypotheses and are removed.

In addition to the removal of unpromising dips, these syllable boundaries are used to cut off the residual path for optimal state sequence search inside a syllable as **Figure 5** shows. For example,

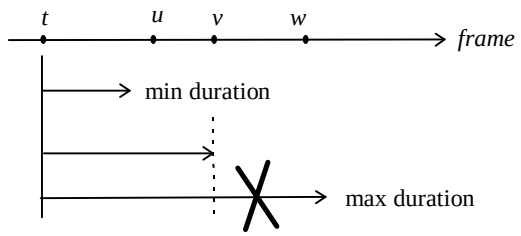

Figure 5: Use of syllable boundary to cut off residual path

if frame *v* in **Figure 5** is a reliable syllable boundary hypothesis, the Viterbi path starting from *t* is cut at frame *v*, instead of frame *w* as is required by the concatenated syllable matching algorithm [2]. The computation thus saved would be significant if the maximum syllable duration constraint is set too long, which is often the case for most of the syllables.

4. USE OF PHRASE BOUNDARY

Since the presence of a phrase boundary is often marked by phrase-final lengthening [7], the duration of each recognized syllable is investigated to check for possible phrase boundaries. If the duration of the recognized syllable is, say 2, standard deviations away from its mean value, this syllable is presumed to be a phrase final, and the score accumulation for words crossing this syllable is stopped. To view it in another way, if a node (character) in the tree is presumed to be a phrase-final, all its children nodes will not be traversed; instead, a compulsory word transition will occur at this phrase-final node. For example, if node C in **Figure 6** is a phrase final, words terminated at node E and F will be invalidated thus are discarded without completing

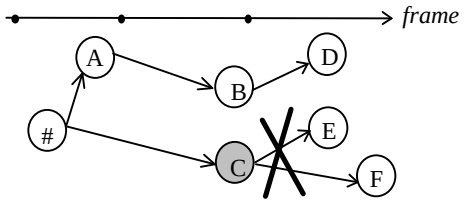

Figure 6: Use of phrase boundary to prune word candidates

the scores. On the other hand, since node B is not a phrase final, the score accumulation of the words terminated at node D will continue. Though this may seem hazardous because words are pruned prematurely before complete scores are known, it is not detrimental. Even correct word is chopped mistakenly due to a false phrase boundary hypothesis, by taking every character as a monosyllabic word in our lexicon, it will be promisingly recovered by the language model later on.

5. EXPERIMENTAL RESULTS

5.1. Speech Database

The speech database is recorded using a noise-canceling headphone with 16kHz sampling frequency in a quiet office environment. On these digitized speech the LPC analysis is performed to prepare appropriate feature vectors for the acoustic modeling of 416 base syllables and 5 lexical tones; i.e., 149 intra-syllable right-context-dependent (RCD) phone-like units for base syllable [1] and 23 context-dependent models for tone [2]. Speech of isolated syllables and phonetically balanced sentences from 40 male speakers is used to train the speaker-independent acoustic model. Results presented in this section are averaged by 3 serious male speakers reading articles excerpted from local newspapers of about 1,500 characters for each speaker. The language model applied is the word class bigram based on starting/ending characters as mentioned in section 2.2.

5.2. Integration of Acoustic and Linguistic Knowledge

A two-staged recognition architecture [2] is taken as the baseline system in this experiment. Compared to the baseline system, a error reduction rate of 43.5% is observed when acoustic and linguistic knowledge is integrated in a single stage as **Table 1** lists. Although the two-staged architecture can attain satisfactory performance [2] when the acoustic model is reliable, this is not the case for poor acoustic model as this experiment shows, where the importance of the integrated architecture is justified.

	Two-staged	Integrated
Character Accuracy	61.63%	78.33%

Table 1: Average character accuracy for the 3 male speakers with speaker independent acoustic model.

5.3. Use of Boundary Conjectures

The results of using prosodic information for syllable and phrase boundary conjectures are graphed in **Figure 7**. About 36% time save is achieved with only 4.85% error rate increase when the probability of voicing is used to hypothesize the syllable boundaries. The speedup can also be explicated comparing the number of syllable boundary hypotheses in these two cases. For the case when syllable boundary hypotheses are conjectured via LPC gain dips only, the number is 4.43 hypotheses per actual syllable boundary; yet the number reduces to 1.49 hypotheses per actual syllable boundary when the probability of voicing is used to prune unpromising syllable boundary hypotheses- 66.37% of the

boundary hypotheses by LPC gain dips are removed.

If the phrase boundary conjectured via phrase-final lengthening is also used, another 17.5% speedup is observed in this experiments, with only, as expected, 0.79% error rate increase. Undeniably, the result is the art of tradeoff: the more conservative we set the duration threshold hence the less the accuracy degrades, the less obvious the speedup is observed.

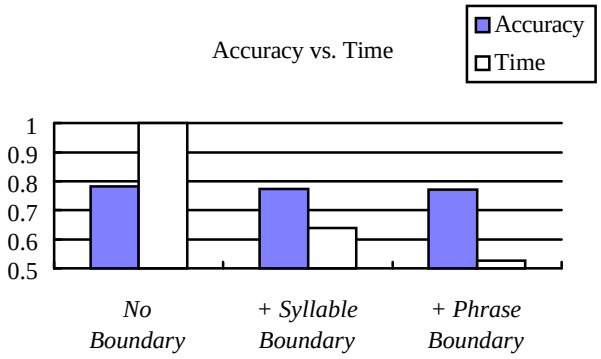

Figure 7: Comparison of character accuracy (x100%) vs. relative recognition time (*No Boundary* as 1) using prosodic information; *No Boundary* stands for the case when syllable boundary hypotheses are conjectured via LPC gain dips only.

6. CONCLUSIONS AND FUTURE WORKS

The use of prosodic information to integrate acoustic and linguistic knowledge in continuous Mandarin speech recognition with very large vocabulary is shown in this paper to be effective for complexity reduction but with only minimal accuracy degradation. Though phrase boundary hypotheses conjectured via phrase-final lengthening *do* help prune unpromising words, it is beneficial if more prosodic information, like F0 declination, can be used in collaboration for more efficient and reliable phrase boundary hypotheses. To further reduce the complexity incurred from the unification of search space, algorithms for fast match or beam search are desirable.

7. REFERENCES

1. Ren-yuan Lyu, et al., “Golden Mandarin (III) – A User Adaptive Prosodic-Segment-Based Mandarin Dictation Machine for Chinese Language with Very Large Vocabulary”, ICASSP, Detroit, U.S., 1995, pp. 57-60
2. Hsin-min Wang, et al., “Complete Recognition of Continuous Mandarin Speech for Chinese Language with Very Large Vocabulary but Limited Training Data”, ICASSP, Detroit, U.S., 1995, pp. 61-64
3. Tai-hsuan Ho, et al., “Fast and Accurate Continuous Speech Recognition for Chinese Language with Very Large Vocabulary”, EUROSPEECH, Madrid, Spain, 1995, pp. 211-214
4. H. Hey, et al., “An Overview of the Philips Research System for Large Vocabulary Continuous Speech Recognition”, International Journal of Pattern Recognition and Artificial Intelligence, Vol. 8, No. 1, 1994, pp. 33-70
5. Yen-ju Yang, et al., “An Intelligent and Efficient Word-Class-Based Chinese Language Model for Mandarin Speech Recognition with Very Large Vocabulary”, ICSLP, Yokohama, Japan, 1994, pp. 1371-1374
6. Simon Haykin, “Neural Networks – A Comprehensive Foundation”, Macmillan College Publishing Co., 1994
7. Entropic Research Laboratory Inc., “Entropic Signal Processing System”
8. Donia R. Scott, “Duration as a Cue to the Perception of a Phrase Boundary”, Journal of the Acoustical Society of America, Vol. 71, No. 4, April 1982, pp. 996-1007