APPLICATION OF SIMULTANEOUS DECODING ALGORITHMS TO AUTOMATIC TRANSCRIPTION OF KNOWN AND UNKNOWN WORDS

Jianxiong Wu and Vishwa Gupta

Nortel, 16 Place du Commerce, Nuns Island, Verdun, Quebec, Canada, H3E 1H6
e-mail: jimwu@nortel.ca, vishwa@nortel.ca

ABSTRACT

This paper proposes simultaneous decoding using multiple utterances to derive one or more allophonic transcriptions for each word. Three possible simultaneous decoding algorithms, namely the N-best-based algorithm, the forward-backward-based algorithm and the word-network-based algorithm, are outlined. The proposed word-network-based algorithm can incrementally decode a transcription from any number of training utterances. Speech recognition experiments for both known and unknown word vocabularies show up to 16% reduction in word error rate when simultaneously decoded allophonic transcriptions are added to the recognition dictionaries. This result holds even for dictionaries originally transcribed by expert phoneticians.

1. INTRODUCTION

One way to improve speech recognition accuracy is to optimize pronunciation dictionary [1]. A data-driven approach is preferred to a knowledge-based or human-transcription-based approach for many reasons. Manual transcriptions are time consuming and in general very expensive. The words have to be re-transcribed every time we move from one dialect to another. Theoretically, it should be possible to accurately transcribe words spoken in a new dialect by data-driven approach using a few sample utterances of the word from that dialect.

One important issue in data-driven pronunciation modeling is how to transcribe words or phrases as a sequence of allophones that best represent the spoken utterances. Given a set of allophone acoustic models, the sequence of allophones that best represent one utterance can be easily obtained by performing a continuous allophone recognition for that utterance with biphone or triphone constraints. However, the resulting allophone sequence obtained from one utterance is frequently erroneous and the use of such an erroneous allophone sequence leads to low recognition accuracy. In order to obtain consistently high-quality transcriptions, we require multiple utterances of the word or phrase to derive one transcription.

Instead of recognizing each utterance as an allophone sequence independently, maximum joint likelihood approaches have been proposed [2][3] to get one transcription from several utterances. The idea is to use the *a priori* information that these utterances correspond to the same allophone sequence. In [2], the best transcription is selected from a set of transcription candidates which are generated by decoding each utterance individually. In [3], the optimal *baseform* transcription is obtained by recognizing input utterances simultaneously. In our experiments we see that the data-driven transcriptions represent closely the regional dialects when we use several utterances from that region to generate these transcriptions.

We discuss several implementations of the simultaneous decoding algorithms. In addition to the N-best and the forward-backward based algorithms, a word-network-based simultaneous decoding algorithm is also presented. This novel algorithm incrementally decodes a transcription from any number of utterances of the word or phrase. These transcriptions are then used in various speech recognition experiments. Up to 16% reduction in word error rate is obtained when simultaneously decoded allophonic transcriptions are added to the recognition dictionaries.

2. SIMULTANEOUS DECODING ALGORITHMS

The goal of a simultaneous decoding algorithm is to find one optimal allophone sequence W^* for all input utterances $U_1, U_2, \cdots U_n$. According to the *Bayes* criterion, W^* should be computed as

$$\begin{aligned} W^* &= \operatorname{argmax}_W \ p\{W|U_1, U_2, ..., U_n\} \\ &= \operatorname{argmax}_W \ p\{U_1, U_2, ..., U_n|W\}p\{W\} \\ &= \operatorname{argmax}_W \ p\{U_1|W\}p\{U_2|W\}\cdots p\{U_n|W\}p\{W\} \end{aligned}$$

Several methods can be used to search for a solution for the above optimization problem. Three methods are discussed in the following sections.

2.1. The N-best based algorithm

Probably the simplest method to perform simultaneous decoding is to use the N-best based approach. In this algorithm, an N-best search algorithm [4][5] is used to generate an individual N-best list for each input utterance independently. These individual N-best lists are merged and rescored using all the input utterances [2]. The transcriptions are then re-ordered based on their joint likelihoods. However, our experience is that this solution is sub-optimal unless N is very large.

2.2. The forward-backward algorithm

The tree-trellis search algorithm [5] can be easily modified to perform simultaneous decoding for multiple input utterances (the same algorithm was used in [3]). The basic idea is to perform a forward *Viterbi* beam search for each utterance independently, then apply a combined backward A^* search [6] for all the utterances simultaneously.

In the forward searches, scores for optimal partial paths from the beginning node to each within-beam grammar node are stored at each frame for each utterance. These forward scores are then used as the heuristics in evaluating incomplete paths for each utterance in the backward search. In the combined backward search, a theory is evaluated by summing the log likelihood of evaluating the same theory for each utterance independently. Since A^* search extends the partial theory with the highest evaluation score, each partial theory extension in the combined backward search is optimal for all utterances.

The algorithm is admissible but not exact since the heuristics used in evaluating incomplete paths in the A^* search is the upper bound on the score of actually extending a common partial theory to the beginning node of the search grammar for all utterances. The major drawback of this algorithm is that it needs to store forward scores for all input utterances. When the number of utterances to be simultaneously decoded is large, the memory requirement may not be manageable.

2.3. The word-network-based algorithm

In order to perform simultaneous decoding for large number of input utterances, a word-network-based algorithm is developed. The algorithm involves four steps:

1. Create an allophonic network for the word or phrase to be transcribed. To keep the algorithm computationally simple, we create an allophonic graph with n+1 nodes, where n is the maximum possible phonemes in any possible transcription of the word or phrase.
2. Score the allophonic network for each utterance independently. Each arc in the allophonic network is associated with a score corresponding to the highest scoring complete path passing through this arc. There are a few efficient ways of scoring this network. For example, the arcs can be scored by a graph search algorithm [7] or by the search algorithm outlined in [8].
3. Merge these individually scored allophonic networks to form an allophonic network scored jointly from all the utterances. The arc scores in the jointly scored allophonic network are the sum of the log likelihoods of the corresponding arc scores in the individual allophonic networks.
4. Find an optimal path through the combined allophonic network. The optimal path is the complete path with the highest summed log likelihood through the arcs in the path.

This algorithm is computationally very efficient since maintaining a combined allophone network requires only a small amount of memory and constructing an optimal path through the combined allophonic network is very fast. Furthermore, The combined allophonic network can be constructed incrementally. That is, once the individual allophonic network for the i^{th} utterance U_i is constructed, the combined allophonic network can be updated and the simultaneously decoded allophonic sequence for utterances $U_1, U_2, \cdots, U_i$ can be obtained. This combined allophonic network can possibly be used to constrain the search space of the next input utterance U_{i+1}.

3. EXPERIMENTS

3.1. Transcription generation for unknown words

In this experiment, the vocabulary of the speech recognition system was specified by the user in the enrollment process. The recognition system had no knowledge about the orthography of user-defined words. Two enrollment utterances were used to define a dictionary entry.

Experimental data was collected from 28 households in a 5 month period through public telephone network. In each household, dictionary entries were enrolled and used by several users. Various speech and non-speech background noises were observed in both enrollment and recognition utterances. In order to make the recognition task artificially difficult, we incorporated a lot of mismatch conditions. Different kinds of telephone sets (including normal phones, speaker phones and cord-less phones) were used in the data collection even in the same household. Users might use one telephone set(s) for enrollment and other telephone set(s) for recognition. Entries enrolled by one user might also be used by other users. Enrollment environment might be different from the recognition environment.

The data set contains 10,132 same-speaker test utterances (i.e., utterances were spoken by speakers who enrolled the corresponding dictionary entries) and 1,104 cross-speaker test tokens. The average dictionary size of these households is 17. The acoustic model contains 373 English allophones. MFCC features (7 cep, 7 delta cep and delta energy) were used in these experiments. Frame synchronous cepstrum mean subtraction was applied in the front-end processing.

Table 1 shows the recognition rates for various systems using different methods to obtain the transcriptions for user-defined dictionary entries.

system	same speaker	cross speaker
I: top 1	90.6%	80.7%
II: top 2	91.0%	81.1%
III: averaging	91.2%	81.2%
IV: simul. decoding	91.1%	83.5%
V: human transcribed	89.9%	84.2%

Table 1: Recognition rates using user-trained dictionary

In the baseline system I, a continuous allophone recognition using fully-connected allophone search network was performed independently for each enrollment utterance. The best allophone sequence from the recognizer was stored in the dictionary. Each dictionary entry thus had two transcriptions. System II used top two choices from the allophone recognizer, so four transcriptions were used to represent each dictionary entry in the system. From Table 1, we can see that using more transcriptions obtained from the two enrollment utterances results in only a small performance gain.

In addition to the two transcriptions used in the baseline system, System III added another transcription generated from an artificial *average* utterance [2] for each dictionary entry. The acoustic features of this artificial utterance are the Gaussian means of the single-density state observation functions of an HMM word model. These Gaussian means were trained using the two enrollment utterances. As can be seen from Table 1, adding this additional transcription results in the same recognition accuracy as using four transcriptions per entry (System II).

System IV also used three transcriptions to represent one entry. However, the third transcription is generated by recognizing two enrollment utterances simultaneously using the forward-backward algorithm. Although the performance improvement for the same-speaker utterances for System IV is the same as the one using either System II or System III, the simultaneously decoded transcriptions in System IV achieved significant cross-speaker performance improvement (14.5% error reduction compared with the baseline system).

For comparison purposes, Table 1 also shows recognition performance of human transcriptions. In System V, two human experts transcribed enrollment utterances into phoneme sequences. Phonemic transcriptions were automatically transformed into allophonic transcriptions according to the decision tree rules from which the allophone units were defined.

As can be see from Table 1, for same-speaker utterances, data-driven transcriptions in systems I-IV consistently out-performed human transcriptions. This is probably due to the fact that data-driven transcriptions are optimized at the allophone level using the same acoustic model as the one used during recognition. Furthermore, transcriptions directly estimated from speech samples seem to capture detailed information on how a speaker-specific pronunciation evolves. For example, we observed that one phoneme in a human transcription was expanded to a sequence of different allophones of the same phoneme in the data-driven transcription. Such speaker-specific pronunciation information improves same-speaker recognition, but degrades cross-speaker recognition. The effectiveness of simultaneously decoded transcriptions in cross-speaker environment is illustrated in Table 1. The performance difference between System IV and System V for the cross-speaker utterances is very small.

Table 2 shows transcriptions of the English name DENNIS generated from two utterances both independently and simultaneously. The last row shows manual transcription for comparison purposes. In the transcriptions, non-numerical characters represent phoneme labels, while the numbers following the phoneme labels are the allophone indices.

utterance 1	[d6ʌ0n17ʌ2s4h4]
utterance 2	[d6æ2d13ɛ4s3s17]
utterances 1 & 2	[d6ʌ0n17ɛ4s4]
human transcription	[d3ɛ0n17ə1s4]

Table 2: Transcriptions of English name DENNIS generated from two utterances independently and simultaneously.

3.2. Transcription generation for known words

The idea here is to see if a dictionary transcribed by phoneticians, when augmented with data-driven transcriptions, will result in improved recognition accuracy. We tested this idea in a speaker-independent speech recognition system. The words in the dictionary were transcribed manually by phoneticians or through letter-to-phoneme rules. All the frequently occuring words were transcribed manually, while some less frequent words were by letter-to-phoneme rules. Additional data-driven transcriptions for the frequent words were added to this dictionary in order to capture regional pronunciation variations.

We used the word-network-based algorithm to generate transcriptions automatically from training utterances as shown in Figure 1.

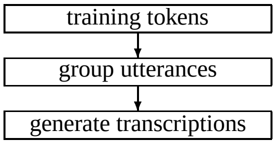

Figure 1: Process of generating automatic transcriptions

Training utterances of a word are clustered into a few groups, generating one transcription per group. The grouping could either be data driven or heuristic. In the experiments mentioned here, either no grouping was performed or the grouping was based heuristically on male/female distinction. All the recognition experiments used gender-independent acoustic models. We imposed a minimum limit of ten utterances for generating one data-driven transcription.

For each orthography, we built an allophonic network for transcription generation as follows:

1. Create a phoneme network from all the phonemic transcriptions of the word (or phrase) in the dictionary. The generated network is a graph with n+1 nodes, where n is the maximum number of phonemes in any transcription. Shorter transcriptions are represented by including null transitions between nodes. The graph overgenerates in most cases. The major constraints this graph imposes are the maximum and minimum number of phonemes in the transcription.
2. For each phoneme arc in the network, create parallel arcs for all phonemes which are in the same confusable class as the current phoneme.
3. Replace each phoneme arc by parallel allophonic arcs corresponding to all possible allophones of the phoneme.

Many different methods may be used to define phoneme confusable classes used in step 2 above. In the following experiments, the confusable classes were based on the phonological features. We used 12 phoneme classes: high front vowel, high back vowel, rounded middle vowel, unrounded middle vowel, low back vowels, glides, liquids, nasal consonants, voiceless fricatives, voiced fricatives, voiceless stops, voiced stops.

Table 3 shows the incremental simultaneously decoded transcriptions from utterances of a French phrase Réno Dépôt. As can be seen from the table, the simultaneously decoded allophone transcription converged to the dictionary transcription (not only the same phoneme sequence but also the same allophone sequence) after 12 utterances. We have observed that for most words or phrases, this convergence is normal. We see a transcription different from that in the dictionary when one allophone of a phoneme is substituted for another allophone of the same phoneme, or when the word or phrase has a frequent regional pronunciation different from those in the dictionary. Table 4 shows a few words for which a new phonemic transcription (different from that in the dictionary) was generated.

Table 5 shows speech recognition accuracy for a recognition task containing 2200 French phrases in the dictionary. Adding automatically generated transcriptions for frequent words reduces error rate by 7% for gender-independent transcriptions, and by 16.7% for gender-dependent transcriptions.

Table 6 reports recognition accuracy for a recognition task containing 1200 English phrases. As in the French recognition

No. of utterances	decoded transcription
1	[r47 œ10 n19 o6 b16 e22 p24 o23]
2	[r47 e10 n19 õ3 b16 e22 p24 o23]
3	[r47 e10 n9 õ9 b16 e22 p24 o23]
4	[r47 e10 n19 õ9 b16 e22 p24 o20]
5	[r47 e10 n19 o16 b16 e22 p24 o23]
6	[r47 e10 n19 o16 b16 e22 p24 o23]
7	[r47 e10 n19 o16 d24 e22 p24 o23]
8	[r47 e10 n19 o16 d24 e22 p24 o23]
9	[r47 e10 n19 o16 d24 e22 p24 o23]
10	[r47 e10 n19 o16 d24 e22 p24 o23]
11	[r47 e10 n19 o16 d24 e22 p24 o23]
12	[r47 e10 n19 o16 d28 e22 p24 o23]
13	[r47 e10 n19 o16 d28 e22 p24 o23]
14	[r47 e10 n19 o16 d28 e22 p24 o23]
15	[r47 e10 n19 o16 d28 e22 p24 o23]
16	[r47 e10 n19 o16 d28 e22 p24 o23]
17	[r47 e10 n19 o16 d28 e22 p24 o23]

Table 3: Incremental simultaneous decoding results for phrase Réno Dépôt. The dictionary transcription is [r47 e10 n19 o16 d28 e22 p24 o23].

phrase	dictionary transcription	new transcription
Gazette Classified	[gəzɛtklæsəfɑjd]	[gəzɛtklɑsəfɑjd]
Mediterraneo	[mɛdətərenio]	[mɛdətərɛnio]
	[mɛdətərænio]	
Le Biftheque	[lebɪftɛk]	[ləbiftɛk]
	[lebiftɛk]	
Royal bank	[rɔjəlbæŋk]	[rwelbæŋk]
Hydro Quebec	[hɑjdrokwibɛk]	[hɑjdrokəbɛk]
	[hɑjdrokebɛk]	

Table 4: Examples of new phonemic transcriptions created by simultaneous decoding algorithm for English phrases. The allophone numbers are left out for simplicity.

task, adding automatically generated gender-independent transcriptions improved recognition accuracy by 9% before model adaptation, and by 13.5% after model adaptation.

4. SUMMARY

Generating allophonic transcriptions from multiple utterances using simultaneous decoding algorithms improves recognition accuracy. Both speaker-dependent and speaker-independent recognition systems benefit from these algorithms. The recognition improves for both known words and unknown words, even for dictionaries transcribed by expert phoneticians. The simultaneous decoding algorithms could possibly be used for transcribing multiple input utterances in other situations where the system has *a priori* knowledge that these utterances contain the same linguistic information.

experiment condition	recognition rate	error red.
original dictionary	85.6%	-
add gender-indep. auto. trans.	86.6%	6.9%
add gender-dep. auto. trans.	88.0%	16.7%

Table 5: French speech recognition results

experiment condition	recognition rate	error red.
original dictionary	72.6%	-
add auto. trans.	75.0%	8.8%
model adaptation	73.8%	4.4%
model adapt. + auto. trans.	76.3%	13.5%

Table 6: English speech recognition results

5. REFERENCES

[1] H. Strik and C. Cucchiarini, (1998) “Modeling Pronunciation Variations for ASR: Overview and Comparison of Methods”, *Proc. of ESCA Workshop on Modeling Pronunciation Variation for ASR*, pp. 137-144.

[2] R. Haeb-Umbach, P. Beyerlein and E. Thelen, (1995), “Automatic Transcription of Unknown Words in A Speech Recognition System”, *Proc. of the IEEE Inter. Conf. on ASSP*, pp. 840-843.

[3] T. Holter and T. Svendsen, (1998) “Maximum Likelihood Modeling of Pronunciation Variation”, *Proc. of ESCA Workshop on Modeling Pronunciation Variation for ASR*, pp. 63-66.

[4] R. Schwartz and Y.L. Chow, (1990) “The N-best Algorithm: An Efficient and Exact Procedure for Finding the N Most Likely Sentence Hypotheses”, *Proc. of the IEEE Inter. Conf. on ASSP*, pp. 81-84.

[5] F.K. Soong and E.-F. Huang, (1991) “A Tree-Trellis Based Fast Search for Finding the N-best Sentence Hypotheses in Continuous Speech Recognition”, *Proc. of the IEEE Inter. Conf. on ASSP*, pp. 705-708.

[6] L.R. Bahl, F. Jelinek and R. Mercer, (1983) “A Maximum likelihood approach to continuous speech recognition”, *IEEE Trans. on PAMI*, Vol. 5(2), pp. 179-190.

[7] V. Gupta, M. Lennig and P. Mermelstein, (1988) “Fast search strategy in a large vocabulary word recognizer”, *J. Acoust. Soc. Am*, Vol. 84(6), pp. 2007-2017.

[8] I.L. Hetherington, M.S. Philips, J.R. Glass and V.W. Zue, (1993) “A^* Word Network Search for Continuous Speech Recognition”, *Proc. of the EuroSpeech*, pp. 1533-1536.