CONTEXT DEPENDENT PHONETIC DURATION MODELS FOR DECODING CONVERSATIONAL SPEECH

Michael D. Monkowski, Michael A. Picheny, and P. Srinivasa Rao

Human Language Technologies Group
IBM Thomas J. Watson Research Center
P. O. Box 704, Yorktown Heights, NY 10598

ABSTRACT

Phonetic context was used to predict the durations of phones using a decision tree. These predictions were used to calculate context dependent HMM transition probabilities for these phone models, which were used to decode telephone conversations from the SwitchBoard corpus. We observed that the duration models do not appreciably improve the word error rate; that more can be gained by modeling phone durations within words than by adjusting for local average speaking rates; and conclude that local or global variations in speaking rate are not major contributors to the observed high error rates for SwitchBoard.

1. BACKGROUND

Conversational speech provides a particularly difficult task for speech recognition. It provides much more variability than either dictation, read speech, or isolated commands. Our efforts to date decoding conversational speech, recorded over the telephone, in the SwitchBoard corpus, have produced word recognition error rates near 50 percent. In an effort to discover the particular causes for this high error rate, we have attempted to model the variations in the duration of phones, and have analyzed their contributions to the difficulty of the problem.

Several groups of researchers [6, 4, 2, 3] have used phonetic duration models to improve the results of speech recognition on other tasks, although they usually had one duration distribution per phone. Pitrelli and Zue [5] have used decision trees to predict phonetic durations, but they did not report recognition results. We therefore sought to improve the performance by using decision trees to model the phonetic duration for the SwitchBoard conversational speech task.

2. CHARACTERIZATION OF DURATIONS

The duration of phones in a Hidden Markov Model (HMM) are modeled by the transition probabilities. In our standard system, we have a set of transition probabilities for each phone. This means that we are tacitly assuming that the phone durations are independent of their context and independent of the speaking rate. It seems reasonable to expect, therefore, that we might improve the error rate by better modeling the effects of context and speaking rate on the phone durations.

We would like, therefore, to estimate transition probabilities that give duration distributions that approximate the actual duration distributions of the phones. If we use the matrix of transition probabilities that we get from training with the forward-backward algorithm, we find that the distribution of durations follows a log normal distribution. That is, the distribution of the logs of the durations follow a normal distribution.

Note that since our HMM topology requires a minimum length of 3, we can never fill in the tail of the lognormal distribution for lengths 1 and 2. We can, however, determine that the distribution is lognormal by looking at the cumulative distribution for lengths 3 and greater.

If we use these HMMs to perform a Viterbi alignment to our data, we find that the distribution of actual durations is also lognormal. It's not clear whether this is because we used Viterbi labels to measure the durations, or because we use Viterbi alignments to assign vectors to classes when making prototypes, or because the phone durations are really lognormal, but irrespective of the cause, it is clear that the HMMs can easily model the shape of the distributions. It is particularly fortunate that only two parameters are needed to define each distribution.

3. CALCULATING CONTEXT DEPENDENT TRANSITION PROBABILITIES

The first step in creating the context dependent models is to use decision trees [5, 1] to classify the log durations based on the phonetic context, which includes five prior phones and five following phones. Then with the means

0-7803-2431-5/95 $4.00 © 1995 IEEE

and variances of the log duration distributions we can generate HMMs for each leaf of the tree. The tree for each phone had, on average, 20 nodes on 5 levels and about 4 leaves.

Our HMMs have three states for each phone. Each state has one forward arc and one self loop giving a simple linear topology. We add skip arcs from the first state to third state, and from the first state to the next phone machine.

In order to calculate the transition probabilities, we used simple heuristics derived from trial and error curve fitting. A useful relation is that the exponential of the mean of the normal distribution of the log duration is also the exponential of the median of the normal distribution, so that it is also equal to the median of the lognormal distribution of the duration. It is also convenient to make the following definition,

> **Definition: The "spread" of a lognormal distribution is the exponential of the standard deviation of the distribution of the logarithms of the data points. We can characterize a lognormal distribution with median (times/divided by) spread.**

We calculated the medians and spreads as a function of the Markov transition probabilities. Looking at these, it was decided that loop probabilities less than 0.2 give distributions that are not very lognormal with more than one forward arc. Also, we tried to get the spread near 1.4 where possible, since this is near the lower end of the range of experimentally determined values for spread from the Viterbi alignments.

median duration min	median duration max	transition probabilities loop (n,n)	forward (n,n+1)	skip (1,3)	skip (1,4)
1.0	1.15	Q	0	0	1-Q
1.15	A	0.2	0	Q-0.2	1-Q
A	B	0.2	R	0.8-R	0
B	5.55	0.5-S	0.5	S	0

where A = 2.3431
B = 3.0814
Q = log(median) / log(A)
R = log(median/A) / 2*log(B/A)
S = 2.06/median - 0.37

Table 1. Heuristics for calculating HMM transition probabilities.

For median durations greater than about 5.5 csec., the probabilities of durations 1 and 2 csec. are negligible, so we can forget about adding the skip arcs and use the relation that the probabilites of the forward transitions are approximately inversely proportional to the medians, to rescale the transition probabilities of the original phone machines. For machines with median durations between 5.5 and 2.3 csec., we add one skip arc, and for machines with median durations less than 2.3 csec., we add both skip arcs. The heuristics used to calculate the transition probabilities are shown above.

4. RESULTS OF DECODING

Using these new HMMs, we decoded 167 SwitchBoard test sentences from the "credit card" topic. Using a "cheating" language model (built on the test data), the word error rate was 29.7 percent without the context dependent models and 28.4 percent word error with the duration models. This is a much smaller improvement than expected.

In order to improve results, we used these new HMMs to create Viterbi alignments; recalculated the durations; and generated new decision trees and HMMs. After four such iterations we tried decoding and got 28.7 percent error.

We then tried training the resulting HMMs using the forward-backward algorithm and got 28.0 percent error. The overall improvement using context dependent models, therefore, was only 1.7 percent.

Using a fair maximum entropy language model, the overall error rate improved from 56.0 to 55.0 percent error; and using a general language model that excluded this topic, the error rate stayed at 56.2 percent word error.

language model	duration models without	duration models with
"cheating"	29.7 %	28.0 %
credit card maximum entropy	56.0	55.1
general, excluding topics	56.2	56.2

Table 2. Decoding results (word error rate) for 167 credit card test sentences

One possible explanation for the failure to see marked improvements with duration modeling may be because of insufficient weighting of the transition probabilities.

We therefore tried adjusting the weights of the HMM transition probabilities relative to the weight of the HMM output probabilities and the weight of the language model. To do this, we raised the output probabilities to a power less than one and then normalized to get total probability one. Then the overall HMM score (log probability) for each sentence was multiplied by a factor greater than one so that we, in effect, raise the weight of the transition probabilities.

We have to do this indirectly, since raising the transition probabilities to a power greater than one and then normalizing would change the length distribution of the HMM.

Our standard processing uses the square root of the output probabilities, so we tried using powers 0.3 and 0.2 to increase the weight. Using the general language model on 40 sentences from the "credit card" topic, the word error rate was reduced from 56 percent at power 0.5 to 54 percent at power 0.3, but then increased to 60 percent at power 0.2.

This increasing error with increasing weight (decreasing output transition power) was characterized by the appearance of many deletions as the silence models ate up frames to the ends of sentences. This is probably due to the unpredictability of silence durations, to our fixed spreads on the duration of all phones including silence, despite the greater actual spread for silence, and to output probabilities for silence that are smoothed due to training to noise frames and to quiet speech frames at the ends of the silence periods.

5. ACCOUNTING FOR THE VARIANCE

These results could have several explanations. Perhaps our decision trees account for very little of the contextual variation or perhaps the variation of duration has little effect on the error rate.

For convenience, let us make the following definition,

> **Definition: "pace" is the ratio of the actual duration to the duration predicted by the decision tree.**

For perfect prediction, the pace would be 1. If the tree predicted 100 percent of the contextual effects, but not the average speaking rate, then the pace would be less than 1 for fast speech and greater than 1 for slow speech.

We can calculate the pace for phones, words, sentences, and so forth, by comparing actual durations to sums of the predicted phone durations. When we look at the distribution of pace values, we find that it also follow a lognormal distribution.

phones	1.59
words	1.47
sentences	1.175
speakers	1.084

Table 3. Spread of the pace, ignoring silence, for the 167 sentence test set

Note that the variation across speakers and across sentences is small compared to the variation across words.

The fraction of the variance accounted for by the context dependent durations can be determined by comparing the variance of the log duration to the variance of the log pace for each phone. For most phones it is 27+/-12 percent of the variance. Weighting the percentages by the frequency of occurrance of the phones, excluding silence, gives 30 percent of the variance. This is comparable to the results obtained by Pitrelli and Zue [5], however they report their results as a percentage of the variation across all phones whereas we are looking at individual phones. Some of the variance was already accounted for by using independent phone models.

Looking at the most common words in the test set, we find that over 50 percent of the variance of the duration for the words "not" and "get" are accounted for by the tree. Others, such as "I" and "Uh" have 10 percent or less accounted for.

not	58	the(02)	27	I	10
of(03)	58	them(01)	26	to(01)	9
get	56	and(03)	23	(silence)	5
don't	41	use(02)	16	a(02)	2
just	41	know	15	Uh	-5
I'm	40	of(04)	14	have(02)	-6
they	32	that(01)	13	(silence)	-7
you(02)	31	it(02)	12	my	-35
and(02)	28				

Table 4. Percent of variance accounted for by the context dependent durations for all word pronunciations with 20 or more occurrances in the test set.

6. PREDICTING THE PACE

If we look at the variance of (log(phone pace) - log(word pace)) we can calculate how much of the variance could be accounted for if we knew the actual word pace. We find that 57 percent is accounted for, meaning 30 percent from the context tree and 27 percent from the word pace. The 43 percent left over within the words can be due to syllable stress, speaker variations, random processes, or to context effects that we miss.

The reality is that we cannot predict the word pace exactly, therefore our estimate of the local speaking rate must restrict us to accounting for less than the 57 percent expected in the ideal case. We already can account for 30 percent, though.

If we estimate the current word's pace by that of the previous word ignoring silence, "Uh", and "I", our 30 percent gets reduced to -3 percent. We lost everyting we had and more. This happens because the correlation between the pace of the current and previous words is only 0.13.

In fact, we find that the correlation between the pace of the current phone and the pace of the previous phone is only 0.15, so it is unlikely that any estimate of the current word pace would be helpful.

7. CONCLUSIONS

We can conclude, therefore, that although the duration models can account for about 30 percent of the variance in the phone duration, they have little effect on reducing the word error rate. This has two implications: first that the HMMs with standard weighting are not adding to the error rate as a result of overconstraining durations (as was seen for the silence models at higher weight), and are therefore not responsible for the high error rates seen with SwitchBoard; and second, that we cannot use duration constraints to appreciably reduce this high error rate.

REFERENCES

[1] L. R. Bahl, P. V. DeSouza, P. Gopalakrishnan, D. Nahamoo, and M. A. Picheny, "Context Dependent Modelling of Phones in Continuous Speech Using Decision Trees," in *Proc. of the Speech and Natural Language Workshop*, (Pacific Grove, CA), June 1990.

[2] H-y. Gu, C-y. Tseng, and L-s. Lee, "Isolated-Utterance Speech Recognition Using Hidden Markov Models with Bounded State Durations," *IEEE Trans. on Signal Processing*, 39(8), pp. 1743-1752, 1991.

[3] M. Jones and P. C. Woodland, "Using Relative Duration in Large Vocabulary Speech Recognition," in *Proc. of the European Conference on Speech Communications and Technology*, pp. 311-314, 1993.

[4] A. Ljolje and S. E. Levinson, "Development of an Acoustic-Phonetic Hidden Markov Model for Continuous Speech Recognition," *IEEE Trans. on Signal Processing*, 39(1), pp. 29-39, 1991.

[5] J. F. Pitrelli and V. W. Zue, "A Hierarchical Model for Phoneme Duration in American English," in *Proc. of the European Conference on Speech Communications and Technology*, pp. 324-327, 1989.

[6] M. J. Russel and A. E. Cook, "Experimental evaluation of Duration Modelling Techniques for Automatic Speech Recognition," *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, pp. 2376-2379, 1987.