ADDING WORD DURATION INFORMATION TO BIGRAM LANGUAGE MODELS

George Doddington[1,2], Aravind Ganapathiraju[3], Joe Picone[3], Yufeng Wu[3]

[1] National Institute of Standards and Technology, [2] SRI International, [3] Mississippi State University

ABSTRACT

Suprasegmental information, while generally thought to play an important role in speech recognition by human listeners, has shown little promise in previous attempts to integrate into ASR systems. This paper outlines an approach that will successfully exploit suprasegmental information by modeling duration within the context of N-gram language modeling. Results show that up to half of the variance in word-level timing can be explained in terms of a simple bigram duration model. These experiments were conducted using the Switchboard corpus of conversational speech over the telephone. The paper also outlines a way of augmenting the N-gram language model with suprasegmental information.

1. INTRODUCTION

Suprasegmental information is generally believed to play an important role in the recognition of speech by human listeners, and there has been a widespread desire and numerous attempts to incorporate this kind of information in ASR systems. (See, for example, R. Gadde, E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tür, "Prosody Modeling for Speech Recognition and Understanding," Presented at the Hub-5 Conversational Speech Recognition (LVCSR) Workshop, Linthicum Heights, Maryland, USA, June 1999.F. Alleva, X. Huang, M. Hwang, and L. Jiang, "Can Continuous Speech Recognizers Handle Isolated Speech?," Speech Communication, vol. 26, pp. 183-189, June 1998.A. Stolcke, E. Shriberg, D. Hakkani-Tür, and G. Tür, "Modeling the prosody of hidden events for improved word recognition," Proceedings of the 6th European Conference on Speech Communication and Technology, Budapest, Hungary, September 1999.A. Stolcke, E. Shriberg, D. Hakkani-Tür, G. Tür, Z. Rivlin, and K. Sonmez, "Combining words and speech prosody for automatic topic segmentation," Proceedings of the DARPA Broadcast News Workshop, pp. 61--64, Herndon, VA, USA, 1999..) Unfortunately, these attempts have been only marginally successful, at best. A typical method of incorporating suprasegmental information has been to add suprasegmental measures to the segmental feature vector, and then to model the segmental and suprasegmental information jointly with a hidden Markov model (HMM). An alternative would be to incorporate the suprasegmental information at a higher level, namely in the language model. This paper attempts to motivate such an approach and discusses how one might go about doing this.

The N-gram is currently the language model most commonly used in ASR. There are several reasons for this, not the least of which is that it provides generally superior performance. It is also relatively easy to implement and train. These characteristics also make the N-gram model suitable as a means of modeling suprasegmental information. And while the N-gram lacks explicit linguistic structure, much of its power lies in an ability to capture semantic information (by means of the strong correlation of meaning with specific word sequences). Therefore, since suprasegmental information is also strongly tied to meaning, the N-gram would seem to be a natural and promising way to capture it.

2. N-GRAM LANGUAGE MODELS INCORPORATING SUPRASEGMENTAL FEATURES

Candidate features include the usual prosodic features of pitch, timing and energy, as represented by the pitch period in milliseconds (*P*), the time interval in seconds (*T*), and the energy (*E*). A logarithmic transformation of these is suggested, in order to produce a more normal statistical distribution of feature values.

The usual N-gram model is simply the probability of a word given the preceding N-1 words:

$$\Pr\left(w_i \,\middle|\, w_{i-1}, w_{i-2}, \mathrm{L}\right)$$

The question is how to represent the candidate prosodic features at the word level. There is no obvious and correct way of doing this. So, in order to begin, let each prosodic feature be represented as a single scalar attribute of each word. For example, P_j could be the average pitch period, averaged over all voiced frames in word *j*; T_j could be the time interval between the end of word *j-1* and the end of word *j*; and E_j could be the rms energy, averaged over word *j*. When these features are added to the model, the model becomes:

$$\Pr\left(F_i \,\middle|\, F_{i-1}, F_{i-2}, \mathrm{L}\right)$$

where

$$F_j = \begin{bmatrix} w_j \\ \log(P_j) \\ \log(T_j) \\ \log(E_j) \end{bmatrix}$$

2.1 WORD DURATIONS AND THE BIGRAM MODEL

To begin exploration of suprasegmentals, we should start with a model that satisfies the following criteria:

1. Suprasegmental features are limited to those considered most valuable.
2. These features can be modeled easily and adequately.
3. The model is simple and takes little time and effort to implement.

Toward this end, word-level duration is chosen as the single feature that best embodies these characteristics. The reasons are:

Word duration is simple and well defined, at least in terms of speech recognition output.

Word duration data are available for training as a side-effect of traditional speech recognition training.

Word duration modeling may be incorporated as a post-processing step on N-best or lattice output. This allows the use of pre-computed acoustic model likelihoods and thus a much simpler and quicker implementation. (Integrating duration information into the acoustic-level search would certainly improve performance. Positive results should be obtainable without such integration, however, if duration information is of sufficiently significant value.)

4. Our initial attempts at incorporating word durations into our model have focused on a rescoring paradigm using a bigram language model. In this case, the feature vector becomes:

$$F_j = \begin{bmatrix} w_j \\ \tau_j \end{bmatrix}$$

where

$\tau_j = \log(t_0 + t_{end}(j) - t_{beg}(j))$,

$t_{beg}(j)$ and $t_{end}(j)$ are the beginning and ending times of word j, respectively, and

t_0 is 10 msec.

The N-gram probability then becomes:

$$\begin{aligned} \Pr(F_i \mid F_{i-1}) &= \Pr(w_i, \tau_i \mid w_{i-1}, \tau_{i-1}) \\ &= \Pr(\tau_i \mid w_i, w_{i-1}, \tau_{i-1}) \Pr(w_i \mid w_{i-1}, \tau_{i-1}) \end{aligned}$$

3. STATISTICAL ANALYSIS ON THE SWITCHBOARD CORPUS

The feasibility of creating and using the model described above is a critical issue. Feasibility hinges on training data and coverage. Two key questions are: How much training data is required to estimate model parameters, and how much coverage is afforded by the fraction of models that are adequately trained? These questions were explored for telephone conversational speech using a large (700k word) subset of the Switchboard corpusJ. Godfrey, E. Holiman and J. McDaniel, "SWITCHBOARD: Telephone Speech Corpus for Research and Development," Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 1, pp. 517--520, San Francisco, CA, USA, March 1992.www.ldc.upenn.edu/readme_files/switchbrd.readme.html, designated here as SBC_{ss}. SBC_{ss} was used for all the experiments discussed in this section.

Figure 1. Coverage and occurrence statistics for the most frequently occurring word bigrams in the SBCss dataset. shows both the coverage and the amount of training data available for the SBC_{ss} dataset. For these data, half of all word tokens are covered by 3000 bigrams, and there are 30 or more occurrences of each of these bigrams in the data subset. (For this analysis, “sentence-begin” and “sentence-end” were excluded from the tabulation.)

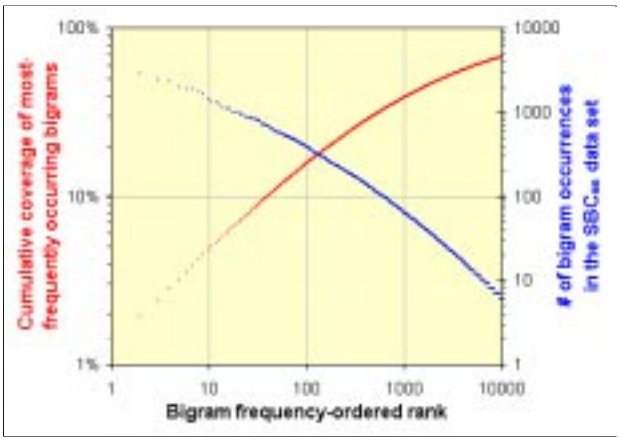

Figure 1. Coverage and occurrence statistics for the most frequently occurring word bigrams in the SBC_{ss} dataset.

From these results, it appears that a duration model may be feasible, because a substantial fraction of all Switchboard speech is covered by bigrams with a substantial amount of training data. The question, then, is whether such a model might be effective. To answer this question, we need to study bigram statistics in detail. Table 1. The most frequently occurring word bigrams in the SBCss dataset. lists some of the most commonly occurring word bigrams. These bigrams represent stereotypical conversational interactions that might be expected to exhibit correspondingly stereotypical and therefore predictable suprasegmental characteristics. It is therefore reasonable to expect the duration of these words to be more predictable and exhibit lower variance when they appear in these specific bigram contexts.

Table 1. The most frequently occurring word bigrams in the SBC_{ss} dataset.

Word 1	Word 2	count	Word 1	Word 2	count
YOU	KNOW	7538	I	THINK	2822
AND	UH	2396	I	DON'T	2350
IN	THE	1978	AND	I	1869
OF	THE	1845	A	LOT	1715
KIND	OF	1612	I	I	1410
LOT	OF	1339	IT	WAS	1289
I	MEAN	1270	DON'T	KNOW	1129
TO	BE	1111	I	GUESS	1103
YEAH	I	1083	DO	YOU	1007

In order for the bigram duration model to contribute to ASR performance, the duration of the words in a bigram must be sensitive to the identity of the bigram. Specifically, the duration of the second word in a bigram must be a function of the bigram and/or the duration of the first word, so that the variance of the duration given this information is much smaller than absent this information. To get a quantitative appreciation for this, duration statistics were tabulated for the SBC_{ss} dataset. Word begin/end times were assigned for all words using an ASR system to perform forced alignment.

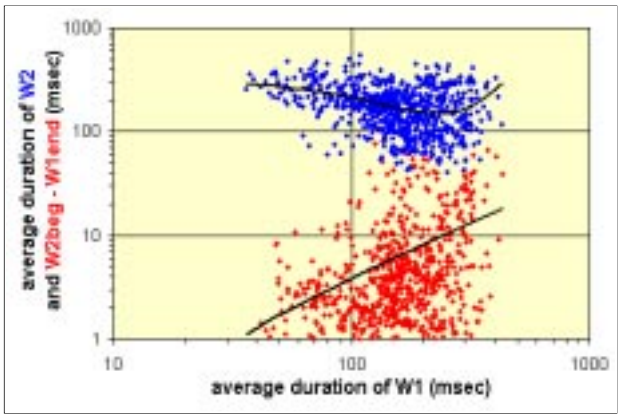

Figure 2. Average duration statistics for the 750 most frequently occurring word bigrams in the SBC_{ss} dataset.

Figure 2. Average duration statistics for the 750 most frequently occurring word bigrams in the SBCss dataset. shows a scatterplot of average duration statistics for the top 750 word bigrams, with the average being computed in the log domain (as defined for the duration features). The average duration of word 2 and the average between-word duration are plotted versus the average duration of word 1. Each of these bigrams occurred more than 80 times in the SBC_{ss} dataset. There appears to be a weak inverse correlation between the duration of word 1 and word 2, at least up to 200 msec. There also is a positive correlation between the inter-word time and the duration of word 1. Note, however, that most between-word average durations are less than 10 msec, which was the frame period used by the ASR word alignment system. (Thus, most of the time there is *no* inter-word gap for these most common bigrams.)

More informative than this global scatterplot would be a plot of the average duration statistics for a *specific* word in a variety of different bigram contexts. This is shown in Figure 3. Average duration statistics for the 750 most frequently occurring word bigrams in the SBCss dataset that include the word "I". for the most common word in the Switchboard corpus, "**I**". This figure is a scatterplot of average durations for all common bigrams that include the word "**I**". Note that the average duration of "**I**" ranges over a factor of 2, depending on the context. The duration statistics of "**I**" in bigram context are illustrated in more detail in Figure 4. Duration distributions for the word "I" in several bigram contexts., which shows the probability distribution function of the duration of "**I**" in several different bigram contexts. Note that for durations of 200 msec or more, there is an order of magnitude difference in duration probability for short contexts such as "**I MEAN**" and "**I GUESS**" as compared to long contexts such as "**I I**" and "**I <END>** ". Note also the (apparently anomalous) high probability of minimum duration, for all contexts. It would be interesting to know whether this is caused by the model-imposed duration constraint, or by misalignment during recognition.

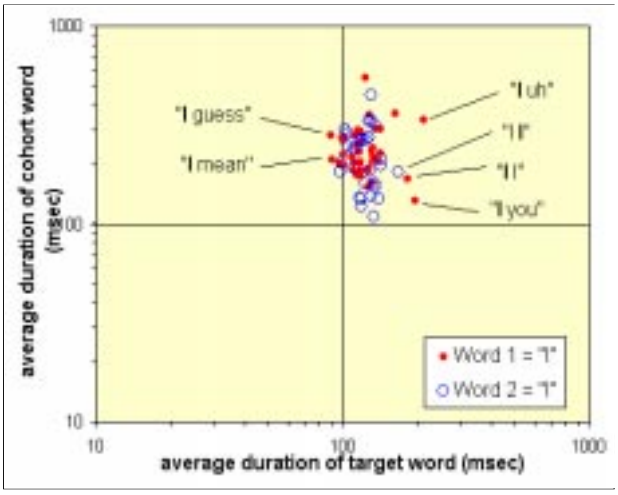

Figure 3. Average duration statistics for the 750 most frequently occurring word bigrams in the SBC_{ss} dataset that include the word "**I**".

In addition to modeling differences in average duration, it is also important that the bigram model be able to predict durations with improved accuracy – i.e., with reduced variance. To study this aspect of the model, a subset of common words was chosen for analysis. To ensure a variety of bigram contexts, choice was limited to those words that occurred as the second word in at least 20 of the top 750 bigrams. There were seven such words, namely "**A**", "**I**", "**IT**", "**THAT**", "**THE**", "**TO**", and "**YOU**". For these words, the average bigram-dependent variance of duration was compared to the grand variance computed over all cohort contexts within the top 750 bigrams. A reduction in variance ranging between 10 and 20 percent was observed for the seven words studied, as listed in the "bigram" row in Table

2. Percent reduction in the variance of the duration of word 2, as a function of the bigram context and linear prediction based on the duration of the first word, for seven frequently occurring words..

In addition to the bigram context, there is also the potential to predict the duration of word 2 based on the duration of word 1. To determine the contribution from this source, covariance statistics for bigram word durations were computed for the top 750 bigrams. This gives, for a simple linear predictive model, a reduction of between 15 and 30 percent, as listed in the "duration" row in Table 2. Percent reduction in the variance of the duration of word 2, as a function of the bigram context and linear prediction based on the duration of the first word, for seven frequently occurring words.. These two factors, when combined, yield a total variance reduction of between 25 and 45 percent, depending on the bigram.

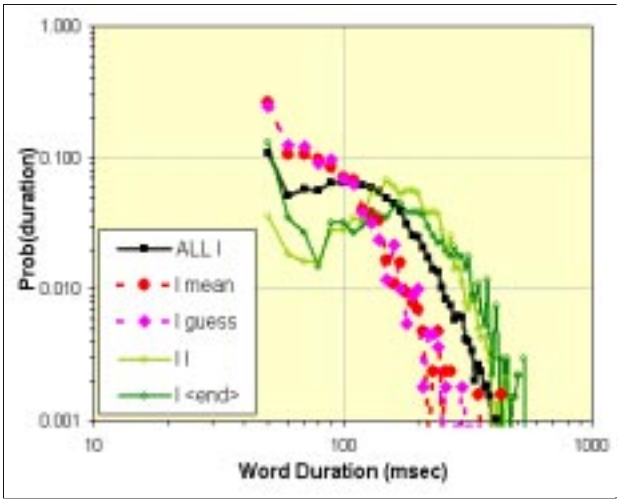

Figure 4. Duration distributions for the word "**I**" in several bigram contexts.

Table 2. Percent reduction in the variance of the duration of word 2, as a function of the bigram context and linear prediction based on the duration of the first word, for seven frequently occurring words.

Word:	**A**	**I**	**IT**	**TH AT**	**THE**	**TO**	**YO U**
bigram	8	9	11	22	8	11	12
duration	30	23	25	27	31	39	16
total	**36**	**30**	**33**	**43**	**37**	**46**	**26**

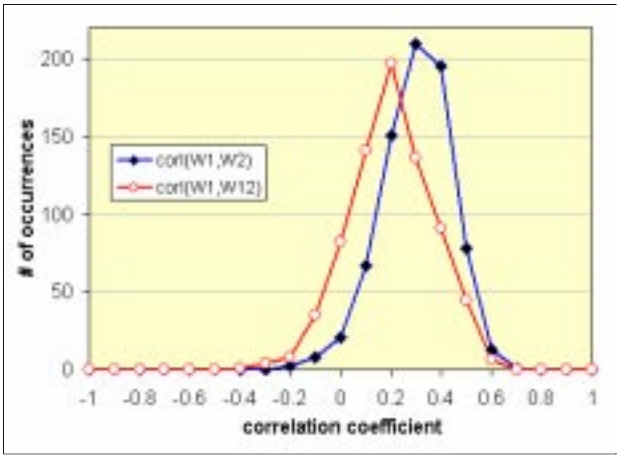

Figure 5. Histogram of correlation coefficients between word durations for the top 750 bigrams. Correlation is computed between

the duration of word 1 and that of both word 2 and the between-word gap.

Figure 5 is a histogram of the correlation between word durations for all 750 bigrams. Note that the between-word correlations average around 0.3. Note also that the first word in the bigram and the duration of the between-word gap tend to be correlated. This suggests the possibility of improving performance by including this between-word gap in the duration model.

4. PLANS

Between now and the workshop in December we plan to implement and test the model that we have outlined. Existing Switchboard models will be used to perform a conventional ASR decoding of a Switchboard test set. Bigram duration models will then be used to refine the search on lattice output from the conventional system. In order to implement this system, we need to know how to use the bigram duration model to calculate language model probabilities, and we need to have a robust method of estimating model parameters.

4.1 COMPUTING LANGUAGE MODEL PROBABILITIES

For ASR decoding, what we actually need to compute is the probability of the word sequence given the data. In this case, we have

$$\begin{aligned}\Pr(w_i|w_{i-1},\tau_i,\tau_{i-1}) &= \Pr(w_i,\tau_i|w_{i-1},\tau_{i-1})/\Pr(\tau_i)\\ &= \Pr(\tau_i|w_i,w_{i-1},\tau_{i-1})\Pr(w_i|w_{i-1},\tau_{i-1})/\Pr(\tau_i)\end{aligned}$$

From here a couple of reasonable simplifications can be made by asserting that the unconditioned probability of word duration, $\Pr(\tau_i)$, is so broad as to be approximately constant and therefore negligible. Also, perhaps due more to expediency rather than reasonableness, we assert that the probability of w_i may be assumed to be more or less independent of the duration of the preceding word, τ_{i-1}. This gives the following simplification:

$$\Pr(w_i|w_{i-1},\tau_i,\tau_{i-1}) \cong C\cdot\Pr(\tau_i|w_i,w_{i-1},\tau_{i-1})\Pr(w_i|w_{i-1})$$

where C is a constant. This gives a simple means of augmenting the conventional language model with a probability model for the duration of a word, given the identity and duration of the previous word.

4.2 ESTIMATING DURATION MODEL PARAMETERS

The question now is how to model the duration and then how to estimate parameters of the model. Training data will be limited, because we are modeling at the (high) level of words, so a parametric approach is virtually mandatory. As an initial probability model, the Gaussian model serves as an easy choice. It is also a reasonable choice, because of the log transformation performed on the duration. The duration model factors into the ratio of two duration probability models, conditioned on bigram:

$$\Pr(\tau_i|w_i,w_{i-1},\tau_{i-1}) = \Pr(\tau_i,\tau_{i-1}|w_i,w_{i-1})/\Pr(\tau_{i-1}|w_i,w_{i-1})$$

Full covariance models will be needed, to capture the correlation between the bigram word durations. Because of the sparsity of training data, a weighting scheme will be necessary in order to protect against estimation errors by combining bigram-specific probability models with word-specific and word-independent models of word duration. One method of weighting is to weight estimates according to the number of training tokens used versus the number needed. For example:

$$\begin{aligned}\Pr(\tau_i|w_i,w_{i-1},\tau_{i-1}) = &\\ &\Omega_{bigram}\cdot\Pr(\tau_i,\tau_{i-1}|w_i,w_{i-1})/\Pr(\tau_{i-1}|w_i,w_{i-1})\\ &+\ \Omega_{word}\cdot\Pr(\tau_i|w_i)\ +\ \Omega_{global}\cdot\Pr(\tau_i)\end{aligned}$$

where

$$\begin{aligned}\Omega_{bigram} &= C\cdot\left(\frac{N_{bigram}}{N_{bigram}+N_{needed}}\right)\\ \Omega_{word} &= C\cdot\left(\frac{N_{word}}{N_{word}+N_{needed}}\right)\cdot\left(\frac{N_{needed}}{N_{bigram}}\right)\\ \Omega_{global} &= C\cdot\left(\frac{N_{global}}{N_{global}+N_{needed}}\right)\cdot\left(\frac{N_{needed}}{N_{bigram}}\right)\cdot\left(\frac{N_{needed}}{N_{word}}\right)\end{aligned}$$

where N_x is the number of training tokens used to estimate model parameters for model x (with N_{needed} being the number needed to ensure satisfactory estimates), and where C is adjusted so that the weights sum to 1. By using such a weighting scheme, it should be possible to

treat all words in a uniform way, with word duration information contributing to the score in proportion to the amount of training data available.

5. REFERENCES

[1] R. Gadde, E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tür, "Prosody Modeling for Speech Recognition and Understanding," Presented at the Hub-5 Conversational Speech Recognition (LVCSR) Workshop, Linthicum Heights, Maryland, USA, June 1999.

[2] F. Alleva, X. Huang, M. Hwang, and L. Jiang, "Can Continuous Speech Recognizers Handle Isolated Speech?," Speech Communication, vol. 26, pp. 183-189, June 1998.

[3] A. Stolcke, E. Shriberg, D. Hakkani-Tür, and G. Tür, "Modeling the prosody of hidden events for improved word recognition," Proceedings of the 6th European Conference on Speech Communication and Technology, Budapest, Hungary, September 1999.

[4] A. Stolcke, E. Shriberg, D. Hakkani-Tür, G. Tür, Z. Rivlin, and K. Sonmez, "Combining words and speech prosody for automatic topic segmentation," Proceedings of the DARPA Broadcast News Workshop, pp. 61--64, Herndon, VA, USA, 1999.

[5] J. Godfrey, E. Holiman and J. McDaniel, "SWITCHBOARD: Telephone Speech Corpus for Research and Development," Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 1, pp. 517--520, San Francisco, CA, USA, March 1992.

[6] www.ldc.upenn.edu/readme_files/switchbrd.readme.html