Representing the Prosodic Context of Words using Gaussian Mixture Models

Shreyas A. Karkhedkar, Nigel G. Ward

Department of Computer Science, The University of Texas at El Paso
sakarkhedkar@miners.utep.edu, nigelward@acm.org

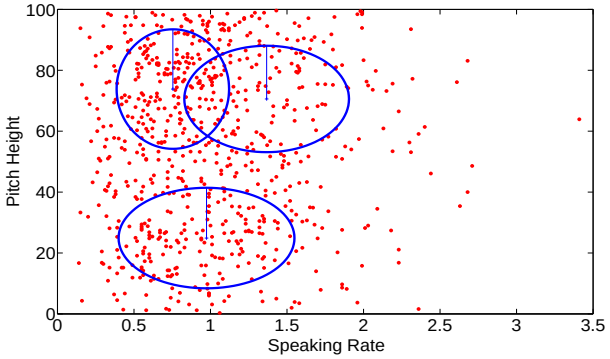

Figure 1: *Distributions and Gaussian models for "see". Details apper in section 1*

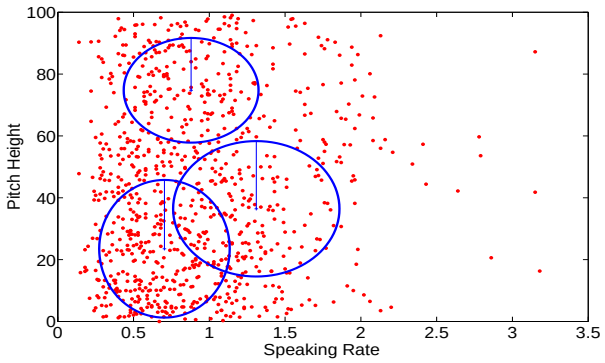

Figure 2: *Distributions and Gaussian models for "because"*

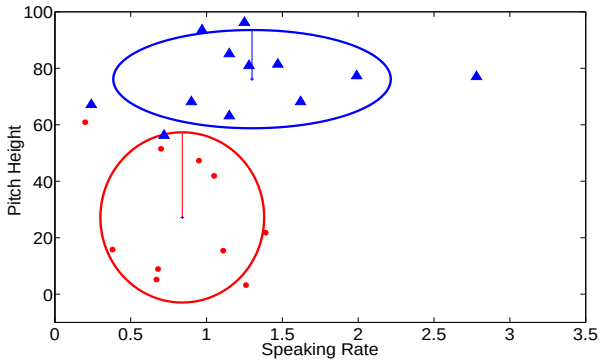

Figure 3: *Distributions and Gaussian models for "corporate" (circles) and "forgotten" (triangles)*

Abstract

Different words tend to appear in different prosodic contexts. For each word, we would like to be able to characterize its typical prosodic contexts of appearance, and, for applications purposes, we would like to do this using probability distributions. This paper describes the phenomena and the need, and explains a way to use Gaussian Mixture Models for modeling prosody.

Index Terms: continuous probability distributions, prosodic features, language modeling, perplexity, lexico-prosodic linkages.

1. Prosodic Contexts of Words

Different words tend to appear in different prosodic contexts. For example, by plotting few hundred occurrences of the words "*see*" and "*because*" (Figures 1 and 2, we see a tendency for "*see*" to occur after regions of higher pitch than "*because*". Similarly, figure 3 shows that the word "*forgotten*" typically occurs after a high pitch region whereas the word "*corporate*" typically occurs after a region that has a lower pitch, and that "*forgotten*" has more variation in its occurrences in terms of previous speaking rate. Further examples of variation in the prosodic contexts of words are given in [1].

Context prosody also can differ among word senses and homonyms, for example, "*may*" and "*May*". Figure 4 shows that "*May*" typically occurs after a regions of somewhat lower pitch. Although a lower pitch height alone would not be enough to disambiguate between the two, when combined with other cues and lexical context this could be informative.

Such linkages between words and their prosodic contexts are not just happenstance, rather they appear to be information that people know and use as clues. Experimental work suggests that words appearing in prosodically unusual contexts are processed more slowly; that is, listeners use prosodic context information as part of the lexical access process [2].

In this paper, Section 2 describes the need to model such regularities of prosodic context, Section 3 critiques the modeling approach of previous work, Section 4 details our approach of modeling prosody as a mixture of Gaussians, Section 5 outlines the method of evaluation and presents the results, and Section 6 presents possible directions for future research.

2. Modeling Prosodic Variations

There are several possible ways to try to understand and model such phenomena. One would be to try to explain them away, as mere surface manifestations of deeper phenomena. Certainly it is true that words have syntactic, semantic, and pragmatic properties, and these properties are in turn associated with prosodic characteristics. It is possible that, eventually, such an explanation could be developed, obviating the need to talk about the

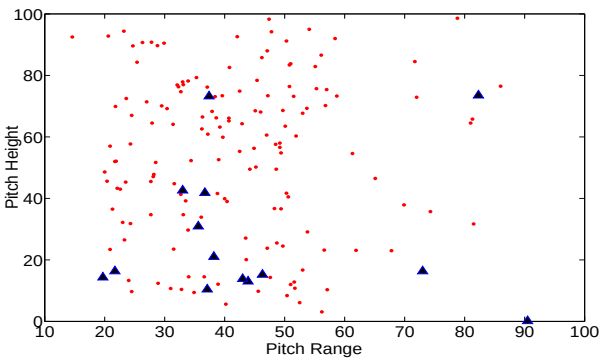

Figure 4: *Distribution for the words "may"(small circles) and "May" (triangles).*

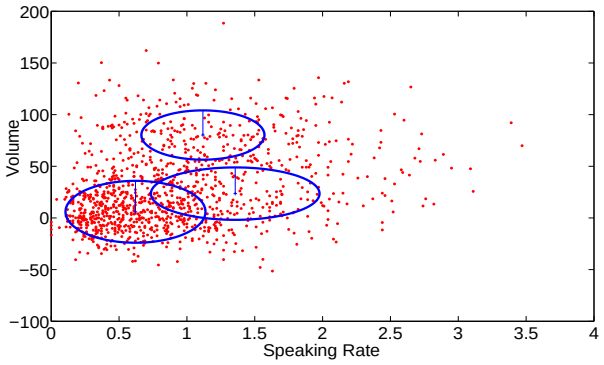

Figure 5: *Distribution of prosodic contexts for the word "right". Negative volume values occur because volume is normalized to make 0 the typical silence value.)*

prosodic contexts of words at all. However it is also possible that these lexico-prosodic linkages are direct, at least in part, that is, that the phenomena reflect direct mental associations between words and their typical prosodic contexts.

Regardless of the underlying psychological reality and the shape of the ultimate model, in the short term, we would like to better model these phenomena, and this is the aim of this paper. Specifically, for any given word, we would like to be able to describe its typical prosodic contexts. Our approach relies on the consideration of the contexts of many occurrences of each word in a large corpus. From this we would like to be able to distill a concise description, for each word, of what prosodic contexts it occurs with; that is, to model the prosodic contexts of each word.

A simple example of such a model would be a listing of the average value, for each of the prosodic dimensions of interest, for each word. However we would like something more descriptive, better to able to capture the range and shape of the distributions of contexts, as seen in the Figures. Such descriptions could be useful in the long-term quest for deeper, mediated explanations of word-context prosody.

They also can help support improved speech applications. For speech synthesis, for example, knowing the typical prosodic contexts for the words of a sentence could help choose prosodic realizations in which each word is in an appropriate context. For spoken behavior analysis, detecting departures from the typical lexico-prosodic linkages can help identify phenomena such as bids for dominance, idiosyncrasies, and emotion [4]. For speech recognizers, knowing which words are likely in a given prosodic context can provide a useful source of information as to which word was actually spoken, especially in cases where the acoustic signal is ambiguous [3].

Such applications need probabilistic models. In particular, for any word of interest and any observed or possible prosodic context, we would like a model to be able to provide an estimate of the likelihood of that word occurring in that prosodic context, and of the likelihood of that prosodic context for that word.

3. Issues in Modeling Variation

For purposes of speech recognition, specifically for the language model, we have previously used a crude way to categorize the prosodic contexts of words [3]. There we discretized each of the prosodic features, namely volume, pitch height, pitch range and speaking rate, into levels. For example, the speaking rate just prior to the occurence of a word could be slow, moderate, or fast. We then counted the frequency of each word in each context. Although crude, this method combined with a basic trigram language model, gave improved estimates of the probabilities of the next word. Measuring the quality of these estimates using perplexity, we found a 4.6% performance improvement.

The current paper is motivated by three major flaws of this discrete model. First, as the underlying prosodic features are continuous, not discrete, categorizing their values leads to a loss of information. For example, a prosodic context where the value for a prosodic feature is slightly less than a category boundary might be give estimates very different from one where the value of the feature is just greater than that category boundary. Thus, the estimates within each level are not at all informed by the information in instances which fall even slightly outside that category.

Second, discretization can lead to data sparsity issues. Dividing each feature into distinct levels comes with a trade-off: the more levels, the more accuracy, but more levels also means that, given a finite set of data to analyze, each will contain fewer observations. For less frequent words, some categories may have no observations at all. Smoothing techniques could be introduced to prevent zero estimates in such cases, but these tend to be ad hoc.

Third, this model has no natural way to capture the distributional properties of polysemous or multifunctional words. For example, Figure 5 shows the prosodic context observations for the word "*right*". When used as an affirmative (*Right!*), we observed that it occurs more frequently towards the start of an utterance and/or after regions with low volume and slow speaking rate. When used as a deictic (*right now*), it is more frequent in the middle of an utterance and with a preceding region of high speaking rate and middle volume. To avoid the issues in semantic and pragmatic analysis, a model of the lexico-prosodic linkages should be able to discover itself, from the data, the various usages and their typical prosodic contexts, if not the actual underlying polysemy structures.

4. Modeling Prosody using GMMs

To address the aforementioned issues with the discrete model, we decided to model prosodic contexts using continuous distributions. Looking at Figure 6, the distribution of occurrences of the word "*I*" with respect to two dimensions of prosody appears to have an oval shape, with occurrences towards the center more

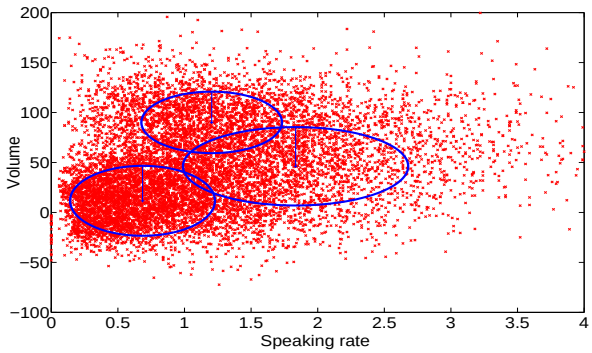

Figure 6: *Distribution of prosodic contexts and Gaussian mixture for the word "I".*

frequent and those toward the edges less frequent. Similar patterns were seen for many other words. Thus we chose to use Gaussian distributions, as a simple distribution that works well for many problems, and appears to fit the patterns we saw.

As the shapes were sometimes lumpy rather than neatly oval, in particular for some polysemous words, we chose to use Gaussian Mixture Models (GMMs), where the probability density function for each word is represented with up to three Gaussian components. For example, the three components found for the word "*I*" are depicted by ellipses in figure 6. The size of the ellipses are relative to their weight. The major and minor axes are relative to the variance in observations.

For this study, we chose to explore these models using the same four prosodic features mentioned above.The features were speaker-normalized [3], but not otherwise adaptive or controlled; and they were computed over fixed-width 300 millisecond windows, rather than being word- or syllable-aligned. Each feature was computed the window ending at the 10 ms interval directly before the onset of the word.

We used standard techniques for determining Gaussian mixture models, following the equations given in [5]. The rest of this section details how we applied them.

First, we changed the constant in equation 1 of [5] to avoid underflow. This did not affect our results since, our evaluation metric, perplexity after combination and normalization, depends only on the likelihoods relative to other words (equation 3).

Second, we separately modeled prosodic contexts which included pitch values and those which did not. In the training set approximately 40% of the data had null readings for at least one of the two pitch-related features (pitch height and pitch range), as many of the words in the corpus were preceded by silence or by unvoiced speech. Therefore, we split the prosodic observations into two disjoint subsets. The first contains prosodic contexts where both pitch height and pitch range values; we call these 4D cases. Figures 1, 2 and 3 represent 2D projections of models obtained from this set. The second set contains all other observations; we call these 2D cases, since here the only usable features are speaking rate and volume. Figures 5 and 6, represent models obtained from the 2D observation subset.

Third, we chose the number of Gaussians to use to model each word depending on the number of instances of that word found in the training data. This is because Gaussian mixture models, as generative models, require many observation instances to accurately model the behavior, and because, in practice, for words with little data, models with too many Gaussians often fail to converge during training. After a little experimentation we set the thresholds for the number of components to use as shown in Table 1. Although these thresholds mean that we can model the prosodic contexts for only a fraction of the total words, the high-frequency words that are covered account for the vast majority of the tokens in the corpus.

Fourth, we constrained our models to have diagonal covariance matrices, to speed training and to allow convergence even with relatively few samples.

Fifth, our GMMs are trained using the standard Expectation-Maximization algorithm. We perform 15 runs for each word, starting with different random initializions. Each run is terminated when the change in log-likelihood (over the previous iteration) falls below 10^{-2} or 100 iterations have been performed. Parameters found on the run with the best log-likelihood on the training data are retained as the model for that word.

Table 1: *Thresholds on the number of observations required for different number of Gaussian components and the distribution of words in the two sets.*

Components	2-D case		4-D case	
	Obs. reqd.	Words	Obs. reqd.	Words
3	100+	246	100+	375
2	50 $\sim$ 99	153	50 $\sim$ 99	269
1	20 $\sim$ 49	406	10 $\sim$ 49	1386
Total		805		2030

5. Experimental Methods and Results

The main goal of the models is to accurately represent information about which prosodic contexts are likely for each word. Once trained, as described above, using one subset of the data, the models are evaluated by seeing how well they describe the prosodic contexts seen for words in a different subset of the data. Of course we do not expect a perfect match, since there is always substantial variation, but if most of the new observations of word w have prosodic contexts which fall within regions represented as highly likely by the model, then the model is a good one.

The training data consists of 985 tracks from the Switchboard [8] corpus of unstructured telephone dialogs, comprising about 80 hours of speech from and 650K word tokens. The tuning data is a disjoint subset of about 35K tokens. This we used as we repeatedly evaluated and refined our method through several preliminary experiments [7]; the configuration which performed best on this subset is the one reported here. The test set contains 16 tracks of speech audio, comprising about 75 minutes of audio and 10441 word tokens. For evaluation purposes, we limited our vocabulary to the top-5000 most frequent words occurring in the training set. All other words are treated as out-of-vocabulary and, hence, excluded from evaluation.

Mathematically, one of the easiest ways to quantify this would be to use the cross-entropy, however we chose to use a less direct method in order to take advantage of our existing software infrastructure. Specifically, we evaluated our model by its ability to improve an existing, standard language model, one based on trigrams [6]. This again reflects a speech recognition scenario, where the recognizer is faced with an acoustically ambiguous signal, and needs to consult the language

model in order to obtain *a priori probabilities* for the various possible words in that context. In order to obtain such probability estimates, our language model used the standard lexical context information (trigrams), but also information provided by our Gaussian Mixture Models. The combined probability estimate was then used to compute the perplexity, which represents, roughly, the amount of uncertainty remaining; thus lower perplexity values are better.

Given a prosodic context, we computed the raw likelihood estimate for each word using that word's Gaussian mixture model and equations 1 and 2. Next, these raw estimates were converted into "scaling factors" using equation 3. Finally, these are combined with the trigram-based probability estimate using equation 4. The parameter "q" in equation 4 was 0.37, the value which maximized performance in the preliminary experiments using the tuning set. The trigram estimates for words which were too infrequent to build Gaussian models were used unchanged. Finally these estimates are normalized to sum to unity.

$$\mathcal{G}_{\check{w}}(\mathbf{x}) = \sum_{k=1}^{K} p_k \mathcal{G}(\mathbf{x}, \mu_k, \Sigma_k) \tag{1}$$

where $\mathcal{G}(\mathbf{x}, \mu_k, \Sigma_k)$ is given by the equation:

$$\mathcal{G}(\mathbf{x}, \mu_k, \Sigma_k) = \frac{1}{\sqrt{2\pi|\Sigma_k|}} e^{-0.5\left(\frac{||x-\mu_k||}{|\Sigma_k|}\right)^2} \tag{2}$$

$$\mathcal{S}(\check{w}) = \frac{N_0 \mathcal{G}_{\check{w}}(\mathbf{x})}{\sum_w \mathcal{G}_w(\mathbf{x})} \tag{3}$$

where N_0 is the number of words having Gaussian mixtures

$$P_{scaled}(\check{w}|\mathbf{x}) = \mathcal{S}(\check{w})^q P_{ngram}(\check{w}) \tag{4}$$

Table 2 shows the performance of the models in terms of perplexity.

Table 2: *Performance of the models*

Model	Perplexity
Baseline n-gram	107.77
Baseline + Gaussians	**105.31**

6. Discussion and Future Work

The perplexity values in table 2 show that GMM-based prosodic models perform better than the baseline trigram model. This shows that prosodic contexts, when modeled as a continuous phenomenon, provide valuable information for predicting the next word.

However, the performance of our model was not better than that obtained by the old, crude discrete prosodic model, with which we saw perplexities below 104 [3]. While the comparison is not exact, because we used different window sizes in that work, overall our GMM models do not seem to be providing additional value. Given that they are also computationally more expensive to generate and evaluate, we are unable to advocate their use for this problem. There are at least possible reasons why they performed less well than expected.

One reason, discovered by examining cases where our GMM models performed poorly, was that the probability estimates they generated could be dominated by a single mismatched dimension, occasionally leading to extremely low probability estimates for words that in fact occurred, whereas the discrete model was more robust to such singular mismatches. Some of these cases could be blamed on the features we used, which were not robust to noise etc., but many were simply due to word appearing in somewhat anomolous contexts [4]. Thus the observed distributions were less tight than we had anticipated, making Gaussians a poor match for the data. Future research might be directed towards developing less sensitive and more robust continuous models, perhaps using some form of fuzzy categorization.

Another reason is that the mathematical justification for using Gaussians for this problem is weak. Variation caused by random noise from diverse uncorrelated factors can be expected to result in distributions that Gaussians model well, for example, if there is an underlying true value that underlies all observations, but is obscured by random measurement errors, then Gaussian distributions are generally appropriate. However, some of the factors affecting the prosodic contexts are partially known, not random and not uncorrelated. Future research might be directed towards understanding the causes of prosodic variation, and in particular developing models which use prosodic features which are better normalized or otherwise correct for known confounding factors.

Finally, our models essentially assume feature independence for the sake of faster computations of model parameters. However, in spoken dialog, prosodic features are often correlated. We are therefore now trying PCA-based dimensionality reduction to address this problem, and also to support models which consider a larger set of prosodic context features.

7. Acknowledgments

We thank Alejandro Vega for helping with the evaluation of the models. This work was supported in part by the NSF as project number IIS-0914868.

8. References

[1] Ward, N. and Vega, A. , "Towards the use of inferred cognitive states in language modeling", in 11^{th} IEEE Workshop on Automatic Speech Recognition and Understanding, 323–326, 2009.

[2] Braun, B., Dainora, A., and Ernestus, M., " An unfamiliar intonation contour slows down on-line speech comprehension," in *Language and Cognitive Processes*, 26 (3), 350-375, 2011.

[3] Ward, Nigel G., Alejandro Vega and Timo Baumann, "Prosodic and Temporal Features for Language Modeling for Dialog", *Speech Communication*, 54, 161-174, 2012.

[4] Ward, N., Vega, A. and Novick, D. "Lexico-prosodic anomalies in dialog", *Speech Prosody*, 2010.

[5] Tomasi, C., "Estimating Gaussian mixture densities with EM - a tutorial", Online: http://www.cs.duke.edu/courses/spring04/cps196.1/handout/EM/tomasiEM.pdf

[6] Stolcke, A. SRILM — An extensible language modeling toolkit, in Proc. Intl. Conf. Spoken Language Processing, 2002.

[7] Karkhedkar, S. and Ward, N., "Using Gaussian mixture models for improved estimates of local prosody-dependent word probabilities", UTEP Department of Computer Science Technical Report, in progress.

[8] Manually corrected Switchboard word alignments. January 29,2003. IISP, Mississippi State University. retrieved from http://www.ece.msstate.edu/research/isip/projects/switchboard