Modeling Systematic Variations in Pronunciation via a Language-Dependent Hidden Speaking Mode

M. Ostendorf, B. Byrne, M. Bacchiani, M. Finke, A. Gunawardana, K. Ross, S. Roweis, E. Shriberg, D. Talkin, A. Waibel, B. Wheatley and T. Zeppenfeld

ABSTRACT

This paper describes the research efforts of the "Hidden Speaking Mode" group participating in the 1996 summer workshop on speech recognition. The goal of this project is to model pronunciation variations that occur in conversational speech in general and, more specifically, to investigate the use of a hidden speaking mode to represent systematic variations that are correlated with the word sequence (e.g. predictable from syntactic structure). This paper describes the theoretical formulation of hidden mode modeling, as well as some results in error analysis, language modeling and pronunciation modeling.

1. Introduction

Spontaneous, conversational speech tends to be much more variable than the careful read speech that much of speech recognition work has focused on in the past, and not surprisingly the recognition accuracy is much lower on spontaneous speech. Pronunciation differences, in particular, represent one important source of variability that is not well accounted for by current recognition systems. For example, the word "because" might be pronounced with a full or a reduced vowel in the initial syllable, or the whole initial syllable might be dropped. Increasing the allowed pronunciation variability of words is needed to handle the reduction phenomena that seem to be a cause of many errors in conversational speech. Unfortunately, as many researchers have noticed, simply increasing the allowable set of pronunciations in all contexts often does not help and may even hurt performance, since the gain of including more pronunciations may be offset by a loss due to increased confusability.

If it is the case that pronunciation changes are systematic, then models can be varied dynamically so as to reduce the added confusability. Thus, the goal of the "Hidden Speaking Mode Group", which participated in the 1996 summer DoD workshop on speech recognition, was to develop a method for allowing pronunciation variations depending on a hidden speaking mode. The speaking "mode" would vary within and across utterances and would reflect speaking "style", e.g. indicating the likelihood of reduced or sloppy speech vs. clear vs. exaggerated speech. By changing the allowed pronunciations as a function of the speaking mode, we can account for systematic variability without the increased confusability associated with a static model.

We focus on capturing variability associated with speaking style, rather than on variability due to dialect or background noise, because of the evidence showing that style has a dramatic impact on recognition performance. In a 1995 study done by SRI, speech recorded over a telephone channel under three conditions showed very different error rates. Spontaneous conversational speech gave an error rate of 52.6%, while the same word sequences read and "acted" (simulated spontaneous speech) by the same speakers led to 28.8% and 37.4% error rates respectively [1]. Since the word sequence and speakers are fixed, the drop in accuracy from read to spontaneous speech must be due to style-related differences.

Speaking style appears to be correlated with the word sequence, and therefore it should be at least somewhat predictable from syntactic and discourse structure. For example, content words (especially nouns) are much more often clearly articulated than function words, which can be reduced to the point of having only a few milliseconds of acoustic evidence. The word sequence "going to" can be reduced to "gonna" when the following word is a verb but not if a noun phrase follows. Old or shared information in the conversation is more likely to be reduced than a new word. Similarly, words at the end of a sub-topic or discourse segment are more likely to be mumbled, while the initial phrase after a topic change will be clearly articulated. Syntactic and discourse structure is of course difficult to extract, but simple text analyses may still be useful for predicting speaking mode.

Because text analysis will necessarily be simplistic and it may be based on errorful data from a recognizer, it is important to also rely on acoustic cues to speaking style. It has been well-established that higher speaking rates are associated with higher recognition errors [2, 3], and rate is perhaps the best candidate for predicting pronunciation variations. However, we have anecdotally noticed reduction phenomena in regions of low energy and pitch range, where a speaker may be mumbling. Thus, the hidden speaking mode model will be conditioned on both acoustic cues and language cues.

In the remaining sections of the paper, we describe the mathematical framework and recognition and training algorithms developed for hidden mode modeling, followed by a summary of the experimental results obtained at the workshop and a discussion of the open questions raised by this work.

2. Hidden Mode Modeling

2.1. Mathematical Framework

Mathematically, the standard problem of recognizing the word sequence $\mathbf{w} = (w_1, \ldots, w_N)$ given acoustic observations $\mathbf{x} = (x_1, \ldots, x_T)$ can be expressed using conditional distributions as

$$\begin{aligned} \hat{\mathbf{w}} &= \operatorname*{argmax}_{\mathbf{w}} p(\mathbf{w}|\mathbf{x}) \\ &= \operatorname*{argmax}_{\mathbf{w}} p(\mathbf{x}|\mathbf{w})p(\mathbf{w}) \\ &\approx \operatorname*{argmax}_{\mathbf{w},\mathbf{q}} p(\mathbf{x}|\mathbf{q})p(\mathbf{q}|\mathbf{w})p(\mathbf{w}), \end{aligned}$$

where $\mathbf{q}$ is a phone sequence associated with the word sequence, $p(\mathbf{x}|\mathbf{q})$ is the acoustic model, $p(\mathbf{q}|\mathbf{w})$ gives the pronunciation likelihoods, and $p(\mathbf{w})$ is the standard language model. With hidden mode conditioning, these equations become

$$\begin{aligned} \hat{\mathbf{w}} &= \operatorname*{argmax}_{\mathbf{w}} p(\mathbf{w}|\mathbf{x}, \mathbf{y}) \\ &= \operatorname*{argmax}_{\mathbf{w}} \sum_{\mathbf{m}} p(\mathbf{x}|\mathbf{w}, \mathbf{m})p(\mathbf{m}|f(\mathbf{w}), \mathbf{y})p(\mathbf{w}) \\ &\approx \operatorname*{argmax}_{\mathbf{w},\mathbf{q}} \sum_{\mathbf{m}} p(\mathbf{x}|\mathbf{q}, \mathbf{m})p(\mathbf{q}|\mathbf{w}, \mathbf{m})p(\mathbf{m}|f(\mathbf{w}), \mathbf{y})p(\mathbf{w}), \end{aligned}$$

where the new variables introduced – $\mathbf{m}$, $\mathbf{y}$ and $f(\mathbf{w})$ – are sequences of mode labels, acoustic cues to the mode, and language cues to the mode, respectively. With hidden mode conditioning, $p(\mathbf{x}|\mathbf{q}, \mathbf{m})$ is the acoustic model and $p(\mathbf{q}|\mathbf{w}, \mathbf{m})$ is the pronunciation likelihood which is interpolated by the mode likelihood $p(\mathbf{m}|f(\mathbf{w}), \mathbf{y})$. The sequence models are simplified using Markov and conditional independence assumptions as in typical recognition systems, e.g.

$$p(\mathbf{x}|\mathbf{q}, \mathbf{m}) = \prod_{t=1}^{T} p(x_t|q_{i(t)}, m_{j(t)})$$

where $i(t)$ and $j(t)$ indicate the phone and mode state associated with time t.

Mode conditioning can be implemented either directly in the acoustic model $p(x_t|q_i, m_j)$ and/or in a pronunciation probability $p(\mathbf{q}|w_i, m_i)$. Both approaches were explored at the workshop. Direct acoustic model mode conditioning can be incorporated by including the mode as a factor in tree-based distribution clustering, together with questions about neighboring phonetic context. Pronunciation probabilities can incorporate mode conditioning in several ways. The simplest approach is to estimate mode-dependent pronunciation probabilities for each possible pronunciation of each word, but these probabilities will only be robust for the most frequent words. As an alternative, we also investigated using decision trees to predict baseform expansion rule probabilities based on the mode or mode-dependent cues in combination with other factors associated with phonetic context. Decision tree pronunciation prediction as proposed in [4] can also be extended to include mode as a prediction factor.

Another design issue to resolve is what time scale the mode varies on. For example, m might be allowed to change at each frame, syllable, word or utterance. Error analyses from the 1995 workshop [5] show that utterance-level factors are not good predictors of error, which suggests that a mode varying within the utterance would be more useful. To restrict the scope of the effort and simplify implementation in recognition and training, we chose to work with a slowly varying mode, assuming that the mode did not change mid-word.

Assuming a word-level mode, the mode sequence includes one mode value m_i for each word: $\mathbf{m} = \{m_1, \ldots, m_N\}$, where N is the number of words (or hypothesized words) in an utterance. The mode likelihood model assumes conditional independence of modes at each word given the acoustic and language cues:

$$p(\mathbf{m}|f(\mathbf{w}), \mathbf{y}) = \prod_{i=1}^{N} p(m_i|f(\mathbf{w}), \mathbf{y}(w_i)).$$

The distribution $p(m_i|f(\mathbf{w}), \mathbf{y}(w_i))$ is represented using a decision tree with questions about the language cues $f(\mathbf{w})$ and the acoustic cues $\mathbf{y}(w_i)$ within a window of the target word w_i.

2.2. Automatic Training

On a small task, it might be possible to hand-label data with modes according to a coding system developed to capture pronunciation-related speaking style variation. However, our experience was that it was difficult to define such a coding scheme, and impractical to label a sufficiently large amount of data by hand. As a result, the work focused on unsupervised learning of initial speaking mode labels through various clustering techniques using acoustic cues y. Given an initial mode labeling, one can estimate mode-dependent pronunciation and/or acoustic models and conditional mode likelihoods, and then iteratively improve all models jointly using Viterbi-style estimation. The problem of finding the hidden speaking "modes" can be thought of as analogous to finding the modes or component distributions of Gaussian mixtures.

Two clustering methods were explored, both based at least in part on decision trees [6]. In the first approach, decision trees are designed to predict regions of recognition error (due at least in part to the acoustic model) vs. regions where the recognizer output was correct, using Chase's error analysis tool [7]. (Errors due to language modeling alone were omitted from clustering.) The leaves of the resulting tree defined a set of "pre-modes". While the acoustic error regions are certainly correlated with different speaking modes, the resulting clusters will not necessarily reflect systematic pronunciation differences. Therefore, the "pre-mode" clusters were subsequently merged using agglomerative clustering with a distance measure on the pronunciation probability distributions of the 100 most frequent words, weighted by the relative frequency of each word.

In the second approach, regions of pronunciation similarity were clustered directly by using the acoustic cues to the mode as features in decision tree clustering to predict baseform expansion rule probabilities. This approach has the advantage of clustering directly on pronunciation variability, which is the goal of the hidden mode

modeling. However, it is only possible when pronunciation variability can be expressed with a small dimensional vector, as in the roughly 20 rules used in the CMU Janus system.

2.3. Recognition

The recognition algorithm relies on a multi-pass search strategy, which reduces the search space by using standard, static-pronunciation hidden Markov models in a first pass of recognition that results in a word lattice or N-best list. The lattice or N-best list must be annotated with at least hypothesized word and silence times, and ideally also with hypothesized phone labels and times, for use in computing acoustic features y.

In rescoring, the dictionary or at least the relative probability of each entry in the dictionary must vary dynamically throughout the utterance, since the mode can change with each hypothesized word. The combined acoustic/pronunciation model of the i-th hypothesized word w_i is given by either

$$p(\mathbf{x}(w_i)|w_i, f(\mathbf{w}^i), \mathbf{y}(w_i)) \quad (1)$$
$$= \sum_m p(\mathbf{x}(w_i)|w_i, m)p(m|f(\mathbf{w}^i), \mathbf{y}(w_i))$$

using a single pronunciation and mode conditioning directly in the acoustic model, or by

$$p(\mathbf{x}(w_i)|w_i, f(\mathbf{w}^i), \mathbf{y}(w_i)) \quad (2)$$
$$\approx \max_k p(\mathbf{x}(w_i)|\mathbf{q}_k) \sum_m p(\mathbf{q}_k|m)p(m|f(\mathbf{w}^i), \mathbf{y}(w_i))$$

using mode conditioning in the pronunciation model, where $\mathbf{x}(w_i)$ and $\mathbf{y}(w_i)$ are the cepstral and mode acoustic features associated with word w_i given its time markings. The language features $f(\mathbf{w}^i)$ are based on the hypothesized word sequence associated with w_i, which will be different for w_i in different N-best word strings. In other words, the mode likelihood provides the probability of the pronunciation in direct acoustic model mode conditioning and acts as an interpolation factor in pronunciation likelihood mode conditioning.

In summary, the differences between hidden speaking mode utterance rescoring and a standard rescoring procedure are that (1) additional acoustic and text analyses are needed for extracting $\mathbf{y}(w_i)$ and $f(\mathbf{w}^i)$ for each hypothesized word w_i in an utterance, (2) a mode likelihood must be computed for each hypothesized word, and (3) each hypothesized word must point to a different pronunciation weight distribution or weighted collection of dictionary entries.

3. Experimental Results

The focus of the Hidden Speaking Mode Group's effort was on developing appropriate models for each of the terms introduced in equations 2 and 3, including the acoustic model $p(x|q, m)$, the pronunciation model $p(q|m, w)$ and the mode likelihood $p(m|f(\mathbf{w}), \mathbf{y}(w))$, as well as on exploring methods for unsupervised learning of the mode. As for all the workshop groups, the experimental paradigm was conversational speech recognition on the Switchboard task [8]. Results were obtained starting from two different baseline systems: an HTK system developed for the workshop trained on 60 hours of data (gender-independent), and the CMU Janus system trained on 140 hours of data (gender-dependent) [9].

A major thrust of the summer effort was data analysis to determine appropriate acoustic features $\mathbf{y}(w_i)$ and preliminary mode clustering. Over 100 features were studied, with some based on forced alignments given the known word transcription (useful in initial mode clustering only), some based on recognized word and phone labels and times, and some that were purely acoustic. The features included various measures of speaking rate, SNR and/or energy, normalized fundamental frequency, presence and duration of silence, and phone label distance measures between different alignment/recognition alternatives. Speaker gender was included as a control to insure that the normalization techniques were effective and the unsupervised mode clustering did not simply learn gender, and in fact gender was never used. The "goodness" criterion for evaluating features was prediction of acoustic modeling errors, i.e. regions where an incorrectly recognized word string had a higher acoustic model likelihood than the correct word sequence. Analyses of individual features showed that normalization is very important, and the best normalization methods were conversation-level. In decision tree error prediction experiments, using recognition on the training data to define a sufficiently large number of error regions, acoustic cues alone gave almost as good performance as the superset of features that included those based on recognizer hypotheses (cross validation error rates of 25% vs. 24%, respectively, compared 36% chance). The most important features were speaking rate (having two measures was better than one) and presence of silence, but SNR also played a significant role. We anticipate that these features may also be useful for research in confidence scoring.

The fact that the presence (but not duration) of silence is important for predicting acoustic modeling errors raises the question of whether silence should be an acoustic cue or a language cue. In other words, silence could be treated as a word in the language model, just as utterance boundaries now are. The silence "word" has been used successfully in the CMU Janus system [9]. Experiments with the HTK system were conducted using one or two silence "words" plus a segmentation boundary "word," where only silences of duration longer than 80 ms were treated as words and 250ms was used as a cut-off for the case when two silence "words" were used. When combined with a language model based on linguistic segmentations and testing with known linguistic segmentations, the silence "words" degraded performance slightly, from 45.1% to 45.8% word error. Using acoustic segmentations, the results were mixed, with the two silence "words" giving a slight improvement, from 46.6% to 46.4% word error. Because the use of a silence "word" blocks actual word context, it is likely that better results could be obtained by including silence "words" but using them in an extended n-gram framework.

Three strategies were pursued for pronunciation modeling, in part because of the differences in the HTK and Janus systems. Results were obtained for modeling pronunciation variations without mode dependence to provide a baseline. In the Janus system, pronunciation variations were generated using a small set of rules for phenom-

ena such as flapping and vowel reduction. Adding pronunciations reduced the error rate from 39.0% to 38.4%, and using pronunciation probabilities derived from the relative likelihood of rule application further reduced error rate to 37.6%. (The Janus results were reported only on male subset of the standard test set.) Analogous experiments were conducted using the HTK system, but adding pronunciations only for the 100 most frequent words. The new pronunciations and their relative likelihoods were based on the results of the Janus system. Perhaps because the Janus pronunciations are tuned to a different acoustic model, there was no gain in performance when using these in the HTK system: 47.0% error baseline performance compared to 47.5% with unweighted additional pronunciations and 47.1% error with pronunciations weighted by their relative frequency. The third approach to pronunciation modeling used distribution clustering, and the baseline (no mode) experiment involved adding stress and syllable structure information to the inventory of clustering questions. Though no recognition experiments have been completed as yet, the training results showed that these features are important in that they are used early in the clustering process. One might expect syllable position and stress to be associated with the relative strength of articulation of a phone (e.g. the strength of a burst for consonants or the distance from a neutral position for a vowel), but it was interesting to see that these were important factors even before many phonetic contextual effects were accounted for.

Finally, although the mode-dependent pronunciation probability distributions are yet to be evaluated in a recognition system, the initial mode clustering experiments did provide evidence to suggest that pronunciation dynamics are at least somewhat predictable from acoustic cues to speaking mode (specifically, speaking rate and normalized energy measures). Pronunciation differences (e.g. differences in the relative likelihood of the pronunciations /ae n d/ and /ax n/ for "and") were found both by clustering probability distributions of phonological rules based on acoustic features, as well as by clustering pronunciation probability distributions associated with the acoustically-derived pre-mode regions.

4. Conclusions

In summary, we have introduced a new approach for handling speaking style variability in speech recognition based on a hidden speaking mode that controls allowable pronunciation variability. Under the assumption that pronunciation variations are systematically related to the speaking mode, a mode likelihood is predicted from acoustic observations such as speaking rate and relative energy as well as from language cues related to the information status (e.g. new vs. old, content vs. function) of words in the local context. We describe different ways of mode conditioning: in the word pronunciation likelihoods (directly or via baseform expansion rules) and in the acoustic models using distribution clustering. Standard training and recognition algorithms are extended to incorporate mode-dependent modeling.

Data analyses were conducted to identify acoustic cues to the mode, and initial pronunciation clustering experiments demonstrate that modes do influence pronunciation likelihood. However, it remains to be shown that mode-dependent acoustic modeling will improve recognition performance. In addition, it is an open question as to where mode-conditioning will be most effective: in the acoustic model or in the pronunciation likelihood. Because of time limitations, many issues related to mode modeling were not explored in depth, such as the use of textual cues to the mode and assumptions about the form of the mode likelihood model. These are just a few of the questions that the idea of hidden mode modeling will raise, making this a fruitful area for future study.

Acknowledgments

The Hidden Speaking Mode group would like to thank: BBN for data resources, SRI for software, the other WS96 groups for help and collaboration on various fronts, Victor Jimenez for heroic efforts in providing recognition lattices and baseline results, and Lin Chase for help with error analysis.

5. REFERENCES

1. M. Weintraub, K. Taussig, K. Hunicke-Smith, and A. Snodgrass "Effect of Speaking Style on LVCSR Performance," these proceedings.
2. D. Pallett, J. Fiscus, W. Fisher, J. Carofolo, B. Lund, M. Przybocki, "1993 benchmark tests for the ARPA Spoken Language Program," *Proc. ARPA Workshop on Spoken Language Technology,* pp. 15-40, 1994.
3. N. Mirghafori, E. Fosler and N. Morgan, "Towards robustness to fast speech in ASR," *Proc. Int'l. Conf. on Acoust., Speech and Signal Proc.*, vol. 1, pp. 335-338, 1996.
4. M. Riley, "A statistical model for generating pronunciation networks," in *Proc. of the Int. Conf. on Acoust., Speech and Signal Proc.*, vol. II, pp. S11.1-S11.4, 1991.
5. *1995 Language Modeling Summer Research Workshop Technical Reports*, Section 2.7, CLSP Research Notes No. 1, Johns Hopkins University, February 27, 1996.
6. L. Breiman, J. Friedman, R. Olshen and C. Stone, *Classification and Regression Trees*, Wadsworth and Brooks, 1984.
7. L. Chase, R. Rosenfeld and W. Ward, "Error-responsive modifications to speech recognizers: negative n-grams," in *Proc. Int'l. Conf. on Spoken Language Processing*, vol. 2, pp. 827-830, 1994.
8. J. Godfrey, E. Holliman and J. McDaniel, "Switchboard: telephone speech corpus for research and development," *Proc. Int'l. Conf. on Acoust., Speech and Signal Proc.*, vol. 1, pp. 517-520, 1992.
9. M. Finke *et al.*, "Janus-II – Translation of spontaneous conversational speech," presentation at the *Large Vocabulary Speech Recognition – Hub 5 Workshop,* 1996.