PROSODY DEPENDENT SPEECH RECOGNITION ON RADIO NEWS

K. Chen, M. Hasegawa-Johnson and S. Kim

Department of Electrical and Computer Engineering and
Department of Linguistics
University of Illinois at Urbana-Champaign, Urbana, IL 61801
`http://www.ifp.uiuc.edu/speech/`

1. Introduction

Does prosody help word recognition? Humans listening to natural prosody, as opposed to monotone or foreign prosody, are able to understand the content with lower cognitive load and higher accuracy [1]. For automatic Large Vocabulary Continuous Speech Recognition (LVCSR), the answer is not that straightforward. Even though successful word recognition and successful prosody recognition have been demonstrated independently in many academic and commercial applications, no result has been reported in the literature that shows improved word recognition on a large-vocabulary continuous speech recognition task with the help of prosody. In 1997, Kompe [2] presented a theoretical proof stating that prosody can never improve word recognition accuracy unless the recognizer uses prosody dependent models. In this paper, we propose a novel probabilistic framework in which word and phoneme are dependent on prosody in a way that improves word recognition.

We propose the use of prosody-dependent allophones based on the "hidden mode variable" theory of Ostendorf et al [3], but with prosody dependence carefully restricted to a subset of distributions that are known to be most sensitive to prosodic context. Specifically, we propose to model prosody dependence of the phoneme duration probability density functions (PDFs), the acoustic-prosodic observation PDFs and the language model, and to ignore prosody dependence of the acoustic-phonetic observation PDFs. In so doing, we create effective models of the most striking and most often reported prosody-dependent allophonic variation, without significantly increasing the parameter count of the speech recognizer.

2. Prosody dependent modeling

In this section, we describe the probabilistic framework we propose for prosody dependent word and phoneme modeling. The task of prosody dependent speech recognition, given a sequence of observed short-time vectors $O = (o_1, ...o_T)$ of the acoustic features, is to find the sequence of word labels $W = (w_1, ..., w_M)$ and the sequence of prosody labels $P = (p_1, ..., p_M)$ that maximizes the recognition probability:

$$[\hat{W}, \hat{P}] = \arg\max p(O, Q, W, P), \tag{1}$$

where $Q = (q_1, ..., q_L)$ is a sequence of sub-word units, typically allophones dependent on phonetic context. Ostendorf et al. [3] suggested expanding equation (1) as:

$$[\hat{W}, \hat{P}] = \arg\max p(O|Q, H)p(Q, H|W, P)p(W, P), \tag{2}$$

where $H = (h_1, ..., h_L)$ is a sequence of discrete "hidden mode" vectors describing the prosodic states of Q, $p(O|Q, H)$ is the prosody-dependent acoustic model, $p(Q, H|W, P)$ is a prosody-dependent pronunciation model, and $p(W, P)$ is a prosody-dependent language model. Equation (2) proposes that every distinct combination of the state variables q and h should be modeled using a distinct acoustic model. In the most straightforward implementation of (2), a recognizer aware of $|h|$ different prosodic contexts would require $|h|$ times as many trainable parameters as a prosody-independent recognizer. In our experiments, we find that the number of parameters required to directly implement (2) is rarely justified by a proportional increase in recognition accuracy. We therefore propose to model only the most salient and widely reported acoustic effect of prosody: intonational phrase-final lengthening [4] and pitch accents.

In [5], we investigated the lengthening of speech segments in the vicinity of intonational phrase boundaries and shew that the phrase final lengthening can be reliably modelled to improve both the word and boundary recognition accuracy. In this paper, we extended the prosody dependence to another common prosody attribute: pitch accents, by making h a two dimensional prosodic vector: $h = [a, b]$, where a is a binary variable indicating if a phoneme q is strengthened (pitch accented), and b is a binary variable indicating if q is lengthened for being in the phrase final position. The observation vector O is also augmented to include two independent feature streams X and Y: $O = [X, Y]$, where X is the conventional acoustic-phonetic observation (typically cepstral coefficients) and Y is the acoustic-prosodic observation. The phrase-final lengthening is modelled by conditioning the state residency time or "duration" of a phonetic state of a HMM on the prosodic variable b. The pitch accents is modelled by conditioning the acoustic-prosodic observation Y on both the phoneme q and the prosodic variable a. We propose to use a prosody-dependent explicit duration hidden Markov model (EDHMM) in order to precisely model prosody dependent phoneme lengthening. The EDHMM of phoneme q_i under prosodic state b_i consists of a sequence of hidden state variables $S_i = (s_{i1}, ..., s_{iN})$, each of which persists for duration d_{ij}, and each of which produces a length-d_{ij} sequence of observation vectors denoted O_{ij}. If, as we propose, the prosodic variables influence only phoneme duration and pitch, then the probability of observing matrix $O_i = [O_{i1}, \ldots, O_{iN}]$ is

$$\begin{aligned}
&p(O_i|q_i, h_i)\\
&\quad= p(X_i, Y_i|S_i, h_i)p(S_i|q_i, h_i)\\
&\quad= \prod_{j=1}^{N} p(X_{ij}, Y_{ij}|s_{ij}, a_i)p(d_{ij}|s_{ij}, b_i)p(S_i|q_i)\\
&\quad= \prod_{j=1}^{N} p(X_{ij}|s_{ij})p(Y_{ij}|s_{ij}, a_i)p(d_{ij}|s_{ij}, b_i)p(S_i|q_i)
\end{aligned} \tag{3}$$

In this paper, the prosody label p_m in equation (2) takes eight possible values that indicate the relative position of the word w_m in an intonational phrase (phrase initial, phrase medial, phrase final, or a one-word intonational phrase) and its binary prominence level (accented, unaccented). The prosodic variable b_i takes only two possible values, indicating whether the corresponding allophone q_i is phrase final, where the phrase final phonemes are defined as the vowel nuclei and coda consonants in the intonational phrase final syllables. The prosodic variable a_i is also binary, indicating whether the corresponding allophone q_i receives a pitch accent. Note that under these definitions, each phonetic state s_{ij} can have at most four different prosody dependent variations: neutral, accented, lengthened, accented+lengthened. However, the number of parameters of s_{ij} is not increased by four times because all the variations of s_{ij} share the same acoustic-phonetic observation PDFs. Meanwhile, the duration PDFs of those phonetic state variations that have the same level of lengthening are shared, and so do the acoustic-prosodic observation PDFs of those having the same level of prominence.

The pronunciation model $p(Q, H|W, P)$ is implemented using a prosody-dependent dictionary in which the connection between the word level prosody variable p_m and the phoneme level prosody variables a_i and b_i are explicitly encoded. The language model $p(W, P)$ is implemented as a prosody-dependent bigram, i.e.

$$p(W, P) = p(w_1, p_1) \prod_{m=2}^{M} p(w_m, p_m | w_{m-1}, p_{m-1}). \qquad (4)$$

3. The acoustic prosodic feature

The fundamental frequency f_0 is generated using the formant program in Entropic XWAVE with probability of voicing (PV) output at the same time as a confidence measure to the extracted f_0. We avoid pitch doubling and halving errors by eliminating f_0 that falls into the doubling and halving clusters of a 3 mixture Gaussian model whose means of the mixture components are restricted to $1/2\mu$, μ, and 2μ, where μ is the estimated utterance mean f_0. We then normalize f_0 by μ and convert it to log scale:

$$\hat{f}_0 = \max(0, \log(f_0/\mu + 1)). \qquad (5)$$

To eliminate unreliable $\hat{f}_0$ measures, those with PVs smaller than a heuristic threshold are replaced by the linear interpolated values $\tilde{f}_0$ based on the $\hat{f}_0$ that have PVs greater than the threshold.

A multi-layer perceptron (MLP) $g(\cdot)$ is trained to transform the interpolated $\tilde{f}_0$ into a new variable Y whose distribution can better fit the Gaussian assumption of HMM: $Y = g(\tilde{f}_0)$. This MLP is trained under the objective of minimizing the mean square error between the MLP output signal and a teaching signal that is designed based on the pitch accent labels. About 85% pitch accent event prediction accuracy is achieved by this method.

4. Experiments

4.1. Database

All but one of the experiments conducted for this research use the Boston University Radio News Corpus (RNC) because it is one of the largest publicly available speech databases transcribed using the ToBI (Tones and Break Indices) prosodic transcription system. RNC speech files include a combination of original radio broadcasts and laboratory broadcast simulations. ToBI transcriptions are available for five talkers (3 female, 2 male). The training and test data include 301 utterances (3775 words, about 3 hours of speech sampled at 16Khz). 90% of the available utterances were randomly selected as training data, while the remaining 10% were used for testing.

	HMM	**EDHMM**
Phone Corr.(%)	64.82	64.84
Phone Acc.(%)	50.98	51.86

Table 1: Phoneme Recognition experiments on TIMIT.

	HMM			**EDHMM**		
	Corr	Acc	#para	Corr	Acc	#para
PI	14.05	2.38	39000	14.32	2.68	43414
PD	33.74	18.9	39789	33.76	19.62	47053

Table 2: % Allophone recognition correctness and accuracy on prosody dependent allophones, and number of parameters of the allophone models. Both PI and PD contain 166 allophones.

In ToBI, break indices are marked to indicate the degree of decoupling between each pair of words. In order to minimize the size of the prosodic search space, only two levels of breaks are distinguished. Breaks with indices higher than 4 (intonational phrase boundaries) are labelled as B4 and breaks with indices lower than 4 are unmarked. Pitch accents are originally marked in 3 main categories: H*, !H* and L*. In this research, we grouped them into a single class. Boundary tones and phrasal tones are ignored.

4.2. HMMs and the acoustic-phonetic features

In all experiments, a 3-state HMM with no skips is used to model all the prosody-dependent allophones, the acoustic-phonetic observation PDF $p(X_{ij}|s_{ij})$ in equation (3) is modelled as 3-component mixture Gaussians, and the acoustic-prosodic observation PDF $p(Y_{ij}|s_{ij}, a_i)$ is modelled as a single Gaussian. The baseline prosody-independent phoneme set is created by eliminating some of the low-frequency function-word-dependent phonemes in the SPHINX phoneme set [9]. A 32 dimensional feature vector consists of 15 MFCC coefficients, energy, and their delta coefficients is used as acoustic-phonetic observations.

5. Results and discussion

Three major recognition experiments were conducted: a prosody-independent phoneme recognition experiment using the TIMIT database (Table 1), a prosody-dependent phoneme recognition experiment using the Radio News Corpus (Table 2), a prosody-dependent word and prosody recognition experiment using RNC (Table 3).

To compare the performance of EDHMM with standard HMM, we conducted phoneme recognition experiments on the TIMIT database using standard 48 phonemes modelled by HMMs of 3 non-skipping states and 3 mixture Gaussians per state. The phoneme recognition accuracy under no grammar condition is improved by .9%, as shown in table 1.

To measure more precisely the influence of prosodic context on phoneme duration and acoustic-prosodic observation PDFs, we conducted prosody-dependent allophone recognition experiments on the Radio News Corpus. Two sets of allophone models were constructed: a prosody-dependent set PD whose prosodic contexts are differentiated by the duration PDFs and the acoustic-prosodic observation PDFs, as shown in equation (3), and a baseline prosody-independent set PI whose prosodic contexts are logically distinct but physically the same, i.e., the duration PDFs under different prosodic contexts are tied and the acoustic-prosodic observation PDFs are removed. By comparing the prosody-dependent allophone recognition correctness and accuracy of PD and PI models with a null grammar (every allphone sequence equally likely), it is possible to assess the strength of the dependence of duration PDFs and acoustic-prosodic observation PDFs over the prosodic context in the RNC database. Table 2 shows the results of this experiment.

To measure the overall performance of prosody dependent recognition, we conducted word recognition experiments and prosody recognition experiments using two types of Acoustic Models (AM) and two types of bigram Language Models (LM). The two types of acoustic models are PI and PD which have been used in the prosody dependent allophone recognition experiment. The two types of language models are denoted as PI and PD as well. Here, PI denotes a LM that contains only plain words with no prosodic variation; and PD is the LM that has the maximal prosody dependence in which a word can have at most eight different prosody-dependent variations. We found the entropy of the test text dropped from 2.05 bits to 1.72 bits after prosody dependence is implemented in language model, at a cost of increasing number of parameters of language models from 5380 to 14751. This result can be explained by the strong correlation between prosody and syntax. By construction, this database includes many word string repetitions, thus word strings in the training data often re-appear in the test data, and so do the prosodically correlated syntactic structures (the syntactic structures that appear with the same prosody context). This improvement of language modelling is further supported by comparing the PI+PI results with the PI+PD results in Table 3. It shows that with the same acoustic model PI, the language model PD can improve word recognition by about .6% over the language model PI. After switching to a better acoustic model PD, the word recognition can be further improved because the interaction between the prosody-dependent acoustic model and prosody-dependent language model increases the likelihood of the word and phoneme paths that are prosodically plausible. This is supported by Table 3 that shows the word recognition accuracy (WRA) of PD+PD+EDHMM has improved about 1.8% over the baseline system PI+PI+HMM.

Table 3 also shows the pitch accent recognition accuracy from this prosody-dependent recognizer. In this case, we compared the decoded word level accent transcriptions with the reference accent transcriptions. The accuracy of pitch accent recognition increases from 55.41% to 79.65% with prosody dependence modelled. This accent recognition results are not directly comparable to the accent event recognition results in the acoustic-phonetic feature construction in section 3, because it is dependent on word recognition accuracy.

6. Conclusions

In this paper, a prosody dependent speech recognizer that models word and prosody in a unified probabilistic framework is proposed. We find that in the Radio News Corpus the knowledge of intonational phrase boundaries and pitch accents can be utilized to improve both word recognition and prosody recognition. Prosody dependent acoustic modeling combined with prosody-dependent language modeling improves word recognition accuracy by an absolute 1.8% using EDHMMs. The best tradeoff between performance and parameterization is achieved by a prosody-dependent HMM recognizer which increases WRA by 1.6% with only 1.7% parameter increase in acoustic modeling and about 21% parameter increase in language modeling.

	AM	**LM**	**HMM**	**EDHMM**
Word	PI	PI	74.89	75.15
	PI	PD	75.52	75.67
	PD	PD	76.50	76.62
Accent	PI	PI	55.41	55.37
	PI	PD	76.75	76.92
	PD	PD	79.61	79.65

Table 3: % word and accent recognition accuracy using PI and PD acoustic models in combination with PI and PD language models.

7. References

[1] L. Hahn, "Native speakers' reactions to non-native stress in English discourse," Ph.D. thesis, University of Illinois at Urbana-Champaign, 1999.

[2] R. Kompe, "Prosody in speech understanding systems," *Lect. Notes in Artificial Intelligence*, 1307:1-357, 1997.

[3] M. Ostendorf, B. Byrne, M. Fink, A. Gunawardana, K. Ross, S. Roweis, E. Shriberg, D. Talkin, A. Waibel, B. Wheatley, and T. Zeppenfield, "Modeling Systematic Variations in Pronunciation via a Language-Dependent Hidden Speaking Mode," *Report of the CSLU 1996 Summer Workshop*.

[4] C. W. Wightman, S. Shattuck-Hufnagel, M. Ostendorf and P. J. Price, "Segmental durations in the vicinity of prosodic phrase boundaries," *J. Acoust. Soc. Am.*, vol. 91, no. 3, pp 1707-1717, March 1992.

[5] K. Chen, S. Borys, M. Hasegawa-Johnson and J. Cole, "Prosody dependent speech recognition with explicit duration modelling at intonational phrase boundaries," in review.

[6] T. H. Crystal and A. S. House, "Segmental durations in connected-speech signals: Current results," *J. Acoust. Soc. Am.*, vol. 83, no. 4, pp. 1553-1573, April 1988.

[7] S. E. Levinson, "Continuously variable duration hidden Markov models for automatic speech recognition," *Computer, Speech and Lang.*, vol. 1, No. 1, pp. 29-45, 1986.

[8] J. D. Ferguson, "Variable duration models for speech," in *Proc. of the symposium on the Application of Hidden Markov Models to Text and Speech*, Princeton, New Jersey, 1980, pages 143-179.

[9] K. F. Lee, "Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 38, No. 4, pp. 599-609, April 1990.