

PROSODY DEPENDENT SPEECH RECOGNITION WITH EXPLICIT DURATION MODELLING AT INTONATIONAL PHRASE BOUNDARIES

K. Chen, S. Borys, M. Hasegawa-Johnson, and J. Cole

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/>

Abstract

Does prosody help word recognition? 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. The prosody attribute that we investigate in this study is the lengthening of speech segments in the vicinity of intonational phrase boundaries. Explicit Duration Hidden Markov Model (EDHMM) is implemented to provide an accurate phoneme duration model. This study is conducted on Boston University Radio News Corpus with prosodic boundaries marked using ToBI labelling system. We found that lengthening of the phrase final rhymes can be reliably modelled by EDHMM, which significantly improves the prosody dependent acoustic modelling. Conversely, no systematic duration variation is found at phrase initial position. With prosody dependence implemented in the acoustic model, pronunciation model and language model, both word recognition accuracy and boundary recognition accuracy are improved by 1% over systems without prosody dependence.

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 phoneme models.

This study focuses most closely on one of the most often reported and most striking examples of prosody-dependent variation: the lengthening of speech segments in the vicinity of intonational phrase boundaries. The lengthening of speech segments in the vicinity of prosodic boundaries has been reported by many phoneticians. Crystal and House [3] reported that the average durations of vowels preceding pre-pausal word-final consonants are considerably greater than those preceding non-prepausal word-final consonants. Beckman and Edwards [4] found that final lengthening occurring at intonational phrase boundaries is a large effect that is highly consistent across speakers and rates. This result implies that lengthening around

boundaries of intonational phrases and higher prosodic domains can be reliably modelled by a boundary dependent duration model. Wightman [5] discovered that segmental lengthening in the vicinity of prosodic boundaries is mainly restricted to the rhyme (vowel nucleus and any coda consonant) of the syllable preceding the boundary. In addition to these results, Fougeron and Keating [6] found that both initial consonants and final vowels at the edges of prosodic domains have more extreme lingual articulations than they would have in other contexts. This suggests that lengthening might affect the duration of speech segments both preceding and succeeding prosodic boundaries. It is interesting to investigate, from the speech recognition point of view, where exactly the lengthening happens and how much it affects the speech recognition.

We propose the use of prosody-dependent allophones based on the "hidden mode variable" theory of Ostendorf et al [7], but with prosody dependence carefully restricted to a subset of distributions known to be most sensitive to prosodic context. Specifically, we propose to model prosody dependence of the duration model and the language model, and to ignore prosody dependence of the mixture Gaussian 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.

We will first present a probabilistic framework for prosody dependent word and phoneme modelling in section 2. In section 3, we will briefly review the Explicit Duration HMM, its training and decoding algorithms, and the extensions we made. We will then present the experiments and results in section 4 and 5, and give conclusions in section 6.

2. Prosody dependent modelling

In this section, we describe the probabilistic framework we propose for prosody dependent word and phoneme modelling. The task of prosody dependent speech recognition, given a sequence of observed short-time vectors $X = (x_1, \dots, x_T)$ of the acoustic features, is to find the sequence of word labels $W = (w_1, \dots, w_M)$ and the sequence of prosody labels $P = (p_1, \dots, p_M)$ that maximizes the recognition probability:

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

where $Q = (q_1, \dots, q_L)$ is a sequence of sub-word units, typically allophones dependent on phonetic context. Ostendorf et al. [7] suggested expanding equation (1) as:

$$[\hat{W}, \hat{P}] = \arg \max p(X|Q, B)p(Q, B|W, P)p(W, P), \quad (2)$$

where $B = (b_1, \dots, b_L)$ is a sequence of discrete “hidden mode” variables describing the prosodic states of Q , $p(X|Q, B)$ is the prosody-dependent acoustic model, $p(Q, B|W, P)$ 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 b should be modeled using a distinct acoustic model. In the most straightforward implementation of equation 2, a recognizer aware of $|B|$ different prosodic contexts would require $|B|$ times as many trainable parameters as a prosody-independent recognizer. In our experiments, we find that the number of parameters required to directly implement equation 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: phrase-final lengthening.

The state residency time or “duration” of an HMM is an implicit random variable with an exponential distribution [8]. The duration of speech segments is significantly non-exponential, being closer in form to a gamma distribution [3]. In order to precisely model prosody-dependent phoneme lengthening, we propose to use a prosody-dependent explicit duration hidden Markov model (EDHMM). The EDHMM of phoneme q_i under prosodic state b_i consists of a sequence of hidden state variables $S_i = (s_{i1}, \dots, s_{iN})$, each of which persists for duration d_{ij} , and each of which produces a length- d_{ij} sequence of observation vectors denoted X_{ij} . If, as we propose, the prosodic variable influences only phoneme duration, then the probability of observing matrix $X_i = [X_{i1}, \dots, X_{iN}]$ is

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

In this paper, the prosody label p_m in equation 2 takes four possible discrete values that indicate whether the corresponding word w_m is phrase initial, phrase medial, phrase final, or a one-word intonational phrase. The prosodic hidden mode variable b_i takes only three discrete values, indicating whether the corresponding allophone q_i is phrase initial, phrase medial or phrase final. A one-word intonational phrase begins with phrase-initial phonemes, and ends with phrase-final phonemes. The pronunciation model $p(Q, B|W, P)$ is implemented using a prosody-dependent deterministic dictionary: for any given word w_m in prosodic context p_m , there is only one allowed pronunciation (Q, B) . The language model $p(W, P)$ is implemented as a prosody-dependent bigram, i.e.

$$p(W, P) = p(w_1, p_1) \prod_{m=2}^M p(w_m, p_m|w_{m-1}, p_{m-1}) \quad (5)$$

In $p(W, P)$, the words that are likely to appear at boundary locations receive larger relative probability than they do in a prosody independent language model $p(W)$ trained from the same text but with no prosody dependence specified.

3. Explicit duration HMM

3.1. Duration density models

In a standard HMM, the duration of a state is an implicit random variable with an exponential probability density function (PDF). The exponential PDF is known to be a poor representation of the distributions of state durations [3]. Many researchers

have reported that the state transition probabilities have minimal effect on word recognition accuracy when prosody is not considered. In the context of prosody dependent recognition, duration modelling has a direct impact on phoneme recognition accuracy, as we will show in section 5.

Two algorithms have been proposed that explicitly model the residence time or “duration” of a phonetic state by extending the underlying Markov chain to a semi-Markov chain. Ferguson [8] proposed an Estimation Maximization (EM) algorithm to estimate a non-parametric probability mass function (PMF) for the state duration. Levinson [9] proposed the continuously variable duration HMM (CVDHMM) in which the state duration probability is modelled as a continuous gamma density function. Ferguson’s algorithm requires more training data than Levinson’s, but has no prior assumption on the parametric form of the duration density function. In addition, Ferguson’s algorithm only requires $O(NT(N+D))$ operations to train, in contrast to $O(N^2TD^2)$ operations in Levinson’s algorithm, where N is the number of states in the HMM, T is the total number of observations in the example, and D is the maximum allowed state duration. Due to this advantage, Ferguson’s algorithm is chosen to be implemented in our system.

3.2. Training and decoding algorithms

Due to the limitation of space, we can not provide a complete review of Ferguson’s algorithm in this section. Instead, we present the extensions we made that are useful for applying this algorithm in LVCSR.

The decoding algorithm of EDHMM has a form that is slightly different from the standard Viterbi algorithm due to the nature of the semi-Markov chain. In analogy to Ferguson’s derivation of the forward-backward algorithm, define $\delta_t^*(j)$ and $\delta_t(i)$ as follows:

$$\delta_t^*(j) = p(x_1, \dots, x_t, s(t) \text{ enters state } j \text{ at time } t+1) \quad (6)$$

$$\delta_t(i) = p(x_1, \dots, x_t, s(t) \text{ leaves state } i \text{ at time } t) \quad (7)$$

where x_t and $s(t)$ are the observation vector and phonetic state at time t , respectively. These probabilities may be computed recursively:

$$\delta_t^*(j) = \max_i \delta_t(i) a(j|i), \quad (8)$$

$$\delta_t(i) = \max_{\tau} \delta_{t-\tau}^*(i) d(\tau|i) b(x_{t-\tau+1} \dots x_t|i). \quad (9)$$

where $a(j|i)$ is the transition probability from $s(t-1) = i$ to $s(t) = j$, $d(\tau|i)$ is the probability that state i persists for τ frames, and $b(x_{t-\tau+1}, \dots, x_t|i)$ is the probability of observing the specified vectors during residence in state i . The existence of Eq. (9) increases the computation by $(D+N)/N$ times over the standard Viterbi algorithm, provided that all the arguments required in (9) are stored in the memory.

Ferguson’s training algorithm and the decoding algorithm specified above were implemented by modifying the Baum-Welch and Viterbi library functions of the Hidden Markov Toolkit (HTK). Due to the efficiency of Ferguson’s training algorithm, it is practical to train EDHMM on a large speech corpus in a reasonable amount of time. The maximum allowed state duration D is chosen automatically by restricting the minimum probability value of the duration PMF. The Token Passing algorithm in HTK is modified to implement the above semi-Markov Viterbi decoding algorithm.

	Boundary Phones	#Phn	#HMM	#EDHMM
IND	None	65	39065	42093
FV	Final Vowels	89	39170	42713
FC	Final Consonants	91	39240	42824
FVFC	FV+FC	105	39345	43519
IV	Initial Vowels	87	39247	42713
IC	Initial Consonants	83	39219	42784
ICIV	IC+IV	102	39401	43462
ICFV	IC+FV	98	39303	43380
IPFP	ICIV+FVFC	153	39688	44718

Table 1: The prosody-dependent phoneme sets, the number of phonemes, and the number of parameters of the models.

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 2 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.

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.

4.2. Boundary dependent HMMs

In all experiments, a 3-state HMM with no skips is used to model both the boundary phones and non-boundary phones, and the observation PDFs are modelled by 3-component Mixture Gaussians. The baseline prosody-independent phoneme set is created by eliminating some of the low-frequency function-word-dependent phonemes in the SPHINX phoneme set [10]. The feature stream consists of 15 MFCC coefficients, energy, and their delta coefficients. In prosody-dependent experiments, the size of phoneme set differs under different types of prosody dependence. Table 1 lists all the prosody dependent phoneme sets we used.

In our labelling system, symbol B4 is used as prefix or postfix to mark the positions of words in the intonational phrases. A word W is labelled as W_B4, B4_W or B4.W_B4 if it is phrase final, phrase initial, or a one-word intonational phrase, respectively. The prosody dependence can be propagated from word level to the phoneme level through prosody dependent dictionaries. For example, in FVFC, the final vowels and final consonants in a phrase final word W_B4 are appended with the _B4 postfix while other phones in this word remain the same. Similarly, in IPFP, the initial consonants and initial vowels in B4_W or B4.W_B4 are prepended with prefix B4_, and the final vowels and final consonants in W_B4 and B4.W_B4 are appended with postfix _B4. Under these definitions, different types of prosody-dependent transcriptions and dictionaries marking different prosody dependent words and phonemes are created.

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

Table 2: Phoneme Recognition experiments on TIMIT.

	HMM		EDHMM	
	IND	PD	IND	PD
FV	25.70	33.93	26.10	34.36
FC	13.22	27.4	13.61	28.02
FVFC	3.13	24.61	3.77	25.36
IC	34.90	25.53	35.28	25.92
IV	34.95	30.09	37.15	30.77
IVIC	33.15	19.10	33.57	19.71
ICFV	23.88	22.89	24.28	23.20
IPFP	1.71	12.19	2.35	12.91

Table 3: Phoneme Recognition Accuracy with boundary and non-boundary phonemes counted as distinct symbols.

5. Results and discussion

Five different recognition experiments were conducted: a prosody-independent phoneme recognition experiment using the TIMIT database (Table 2), a prosody-dependent phoneme recognition experiment using the Radio News Corpus (Table 3), a prosody-dependent word recognition experiment using RNC (Table 4), and two intonational phrase boundary recognition experiments (Tables 5 and 6).

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 2.

To measure more precisely the relationship between prosodic context and phoneme duration, we conducted prosody-dependent phoneme recognition experiments on the Radio News Corpus. For each of the prosodic context definitions in Table 1, two sets of allophone models were constructed: a prosody-dependent set PD whose prosodic contexts are differentiated only by the duration PDFs, as shown in Equation 4, and a baseline prosody-independent set IND whose prosodic contexts are logically distinct but physically the same. Thus, for example, the phonemes X_B4 and X share observation PDFs under condition PD; under condition IND, the phonemes X_B4 and X share all parameters and are in fact identical. By comparing the prosody-dependent phoneme recognition accuracy (PRA) of PD and IND models with a null grammar (every phoneme sequence equally likely), it is possible to assess the strength of the dependence between phoneme duration and each type of prosodic context in the RNC database. Table 3 shows PRA results of this experiment. Note that figures in different rows are not comparable because they are measured under different allophone sets of different sizes.

Table 3 shows that separate modeling of phrase-final and non-phrase final phoneme durations (FV, FC, and FVFC conditions) significantly improves PRA for both HMMs and EDHMMs. This indicates that the lengthening in phrase final rhymes can be modelled by both HMMs and EDHMMs. Conversely, separate modeling of phrase-initial and non-phrase-initial phoneme durations (IV, IC, and IVIC conditions) degrades PRA, indicating that the HMM and EDHMM are un-

AM	LM	HMM	EDHMM
IND	PI	74.89	75.15
IND	PD	75.60	75.68
FVFC	PI	75.03	75.23
FVFC	PD	75.60	75.85

Table 4: % Word Recognition Accuracy using IND and FVFC models in combination with PI and PD language models.

AM	LM	HMM		EDHMM	
		Corr.	Acc.	Corr.	Acc.
IND	PI	84.55	84.47	84.72	84.43
IND	PD	88.07	85.33	88.07	85.37
FVFC	PI	84.59	84.43	84.72	84.43
FVFC	PD	88.11	85.49	88.25	85.49

Table 5: Phrase Initial Boundary Recognition.

able to learn any systematic dependence of phoneme duration on the distinction between phrase-initial and non-phrase-initial position. It can be concluded from these results that distinct modeling of phrase-final phonemes may improve the precision of an HMM or EDHMM phoneme model.

To measure the overall performance of prosody dependent recognition, we conducted word recognition experiments and boundary recognition experiments using two types of Acoustic Models (AM) and two types of bigram Language Models (LM). The two types of acoustic models we used are IND and FVFC as we have discovered in Table 3 that FVFC encodes the best acoustic prosody dependence. The two types of language models are denoted as PI and PD. Here, PI denotes a LM that is completely prosody independent; and PD is the LM that has the maximal prosody dependence in which all 4 types of words: phrase medial, phrase initial, phrase final and one-word intonational phrase are distinguished. The numbers of parameters are 5380 and 8130 respectively in PI and PD Language Models. As can be seen in Table 4, the word recognition accuracy (WRA) of FVFC+PD+EDHMM has improved about 1% over the baseline system IND+PI+HMM. The improvement brought by acoustic modelling is minimal in this result because the language models dominate effectiveness on this database; by construction, this database includes many word string repetitions, thus word strings in the training data often re-appear in the test data. We would expect the effectiveness of acoustic modelling to be more evident on a larger unbiased database.

Table 5 and Table 6 show two types of boundary recognition results. In phrase initial boundary recognition, we create boundary transcriptions by replacing B4_W and B4_W_B4 with B4 and replacing other words with B0. Similarly in phrase final boundary recognition, boundary transcriptions are created by converting W_B4 and B4_W_B4 to B4 and all other words to B0. Roughly 15% of the word boundaries in this database are intonational phrase boundaries, thus simply setting all word boundaries to be B0 gives a boundary recognition correctness of about 84.5%. Prosody-dependent duration modeling (using the FVFC phoneme set) and prosody-dependent language modeling (using the PD grammar) increase boundary recognition correctness by about 3.5%; because of the increased number of word boundary insertions, boundary recognition accuracy increases by only 1%.

AM	LM	HMM		EDHMM	
		Corr.	Acc.	Corr.	Acc.
IND	IND	84.55	84.47	84.72	84.43
IND	PP	87.86	85.33	88.97	85.53
FVFC	IND	84.59	84.43	84.59	84.47
FVFC	PP	88.15	85.49	88.48	85.62

Table 6: Phrase Final boundary Recognition.

6. Conclusions

In this paper, a prosody dependent speech recognizer that models prosody and word in a unified probabilistic framework is proposed. We find that in the Radio News Corpus, the duration lengthening in phrase final syllable rhymes can be utilized to improve acoustic modelling. Prosody dependent phoneme duration models combined with prosody-dependent bigram language modeling improves both word recognition accuracy and boundary recognition accuracy by 1%. Prosody-dependent duration modeling increases the total parameter count of the recognizer by about 10%; prosody-dependent language modeling increases the total parameter count of the recognizer by 6.2%.

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] 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.
- [4] M. E. Beckman and J. Edwards, "Lengthenings and shortenings and the nature of prosodic constituency," in *Between the grammar and physics of speech: Papers in laboratory phonology I*, J. Kingston and M.E. Beckman (Eds), Cambridge: Cambridge University Press, pp. 152-178, 1990.
- [5] 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.
- [6] C. Fougeron and P. A. Keating, "Articulatory strengthening at edges of prosodic domains," *J. Acoust. Soc. Am.*, vol. 101, no. 6, pp. 3728-3740, June 1997.
- [7] 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*.
- [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] 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.
- [10] 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.