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

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

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

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 duration lengthening effects of the 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 New 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 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? The answer is obviously yes for human listeners. For automatic Large Vocabulary Continuous Speech Recognition (LVCSR), the answer is not that straightforward. Even though successful word recognition and successful prosody recognition has been demonstrated independently in many academic and commercial applications, no result has been reported in literature that shows improved word recognition with the help of prosody. In 1997, Kompe [1] presented a theoretical proof stating that prosody can never improve word recognition accuracy unless the recognizer uses prosody dependent phoneme models. Based on this idea, we propose a novel approach that models word and prosody in a unified probabilistic framework in which word and phoneme HMMs are dependent on prosody. Superior word recognition accuracy over baseline systems with no prosody was obtained by our method. This finally proves that prosody can help word recognition in LVCSR if prosody dependence is modelled at phoneme level.

The prosody attributes that we investigated in this study is the duration lengthening of the speech segments in the vicinity of intonational phrase boundaries. We will first present a probabilistic framework for prosody dependent word and phoneme modelling in section 2. We will then present in section 3 some linguistic evidences to support our motivation of modelling this lengthening effects. In section 4, 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 6. Finally, we will give our conclusion in section 7.

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 $X = (x_1, ...x_T)$ of the acoustic features, is to find the sequence of word models $W = (w_1, ..., w_N)$ and the sequence of prosody models $P = (p_1, ..., p_N)$ that maximizes the recognition probability:

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

where $Q = (q_1, ..., q_L)$ is a sequence of sub-word units, typically allophones dependent on phonetic context, and N is the number of word and prosody symbols. (1) can be expanded as:

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

where $B = (b_1, ..., b_L)$ is a sequence of discrete variables describing the prosodic states of Q. In this paper, B only takes three possible values indicating whether an allophone is in a pre-boundary, a post-boundary or a non-boundary position. We will describe our definition of boundary phone in more detail in section 5.2. When prosody is ignored as in conventional speech recognizers, (2) simply becomes:

$$\hat{W} = \arg\max p(X|Q)p(Q|W)p(W). \quad (3)$$

A prosody dependent allophone q_i with prosody state b_i can be modelled using explicit duration Hidden Markov Model (EDHMM),

$$p(X_i|q_i, b_i) = P(X_i|S_i, b_i)P(S_i|q_i, b_i) \quad (4)$$

$$= \prod_{j=1}^{N_i} p(X_{ij}|s_{ij})p(d_{ij}|s_{ij}, b_i)p(S_i|q_i), \quad (5)$$

where $S_i = (s_{i1}, ..., s_{iN})$ is the quasi-stationary states in q_i; X_{ij} is the partial observation sequence occurred in s_{ij}; d_{ij} is the duration of s_{ij}; and N_i is the number of states in q_i. Note that in (5), the prosody state variable b_i only affects the state duration density $p(d_{ij}|s_{ij}, b_i)$. The reason why we did not condition the observation probability $p(X_{ij}|s_{ij})$ on b_i in this study is because it is not clear how much the boundary condition will affect the distribution of cepstral observation vectors and whether the effect is strong enough to be modelled by HMM.

In this study, P in (2) takes four possible discrete values

that indicates whether a word W is phrase initial, phrase medial, phrase final, or a single-worded intonational phrase. The pronunciation model $P(Q, B|W, P)$ is implemented through prosody dependent dictionaries. In the dictionaries, the multiple pronunciations of a word have equal prior chance to be chosen. With the dictionaries, a prosody dependent word sequence (W, P) can be expanded into a prosody dependent phoneme sequence (Q,B) under different boundary phone definitions. The prosody dependence is also modelled in the language model $p(W, P)$. 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.
It is worth noting that in this framework of prosody dependent modelling, the number of parameters in the prosody dependent systems are not significantly larger than those in the prosody independent systems due to the shared observation probability density functions (PDFs). With some slight adjustment, this framework can be generalized to include the dependence over other prosody variables such as phrasal pitch accents and boundary tones.

3. Lengthening at prosodic boundaries

The lengthening of speech segments in the vicinity of prosodic boundaries has been reported by many phoneticians. Crystal and House [2] 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 [3] 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 boundary dependent duration model. Wightman [4] discovered that segmental lengthening in the vicinity of prosodic boundaries is mainly restricted to the rhyme (vowel nucleus and any coda consonants) of the syllable preceding the boundary. In addition to these results, Fougeron and Keating [5] found that both initial consonants and final vowels at the edges of prosodic domains have more extreme lingual articulations than they would be 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. In order to precisely model the boundary lengthening effect, we implemented speech recognizers that explicitly model the duration probability density.

4. Explicit duration HMM

4.1. Duration density models

In standard HMM, the duration of a state is an implicit random varaible with an exponential probability density function (PDF). This does not provide a correct representation of the temporal structure of state durations. Some researchers found that the state transition probabilities have ignorable effects on word recognition accuracy. This is only partially true when prosody is not considered. In the context of prosody dependent recognition, duration modelling has a direct impact in phoneme recognition accuracy, as we will show in section 6.
Two algorithms has been proposed in history that explicitly model duration of HMM by extending the underlying Markov chain to a semi-Markov chain. Ferguson [6] proposed an Estimation Maximization (EM) algorithm to estimate a non-parametric probability mass function (PMF) for the state duration. Levinson [7] proposed the continuously variable duration HMM (CVDHMM) in which the state duration probability is modelled as a continuous gamma density function. Comparing with Levinson's algorithm, Ferguson's algorithm requires a large amount of training data 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 in training, as 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.

4.2. 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 algorithm Ferguson proposed only include the re-estimation formulas for discrete observation PMF and single Gaussian observation PDF. For mixture Gaussian observation PDF, following equations can be used:

$$\gamma_t^r(j,k) = \gamma_t^r(j) \frac{c_{jk}\mathcal{N}(O_t^r, \mu_{jk}, \Sigma_{jk})}{\sum_{m=1}^{M} c_{jk}\mathcal{N}(O_t^r, \mu_{jk}, \Sigma_{jk})}, \qquad (6)$$

where $\gamma_t^r(j)$ is the posterior probability of state j in utterance r; $\gamma_t^r(j,k)$ is the posterior probability of the k^{th} mixture component; O_t^r is the observation vector at t; and μ_{jk} and Σ_{jk} are the mean and variance of the k^{th} mixture component. The k^{th} mixture weight c_{jk} be re-estimated as:

$$\hat{c}_{jk} = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T_r}\gamma_t^r(j,k)}{\sum_{r=1}^{R}\sum_{t=1}^{T_r}\sum_{k=1}^{M}\gamma_t^r(j,k)}. \qquad (7)$$

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 forward and backward probabilities, the maximum posterior probabilities $\delta_t^*(j)$ and $\delta_t(i)$ can be computed recursively:

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

$$\delta_t(i) = \max_\tau \delta_{t-\tau}^*(i)d(\tau|i)b(O_{t-\tau+1}...O_t|i). \qquad (9)$$

This existence of (9) increase 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.
The above training and decoding algorithms are implemented in the Hidden Markov Toolkit (HTK). Due to the efficiency of the 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.

5. Experiments

5.1. Database

As one of a few databases that is designed for study of prosody, Boston University Radio News Corpus consisting of recordings

	What are the Boundary Phones?	size
IND	No boundary phones	65
FV	Final Vowels	89
FC	Final Consonants	91
FVFC	Final Vowels and Final Consonants	105
IV	Initial Vowels	87
IC	Initial Consonants	83
ICIV	Initial Consonants and Initial Vowels	102
ICFV	Initial Consonants and Final Vowels	98
IPFP	Initial and Final Consonants and Vowels	153

Table 1: The definitions of boundary phones for prosody-dependent analysis.

of broadcast radio news stories is used in our study. The recordings are a combination of original radio broadcasts and laboratory broadcast simulations. Files have been transcribed, segmented, and hand-labelled using ToBI prosodic labelling system. In ToBI, break indices are marked to indicate the degree of decoupling between each pair of words. The intonational phrase boundaries are marked by break index of 4. For simplicity, we only distinguish two level of breaks. Breaks with indices higher than 4 are labelled as B4 and breaks with indices lower than 4 are labelled as B0. The training and test sets consist of 301 utterances (about 2 hours of speech sampled at 16Khz) read by five professional announcers (3 female, 2 male).

5.2. Boundary dependent HMMs

In all the experiments, Context-Independent (CI) HMM with 3 non-skipping states are used to model both the boundary phones and non-boundary phones, and the observation PDFs are modelled by Mixture Gaussian with 3 components. SPHINX phoneme set [8] is adopted to form the baseline prosody-independent set with some of the low frequency phonemes merged. The feature stream consists of 15 MFCC coefficients, energy, their delta coefficients. In prosody-dependent experiments, the size of phoneme set differs under different types of prosody dependence. Table 1 listed 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 single worded phrase (such as "Well", "Initially"). 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 attached 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.

6. Results and discussion

To compare the performance of EDHMM with standard HMM, we conducted phoneme recognition experiments on TIMIT database using standard 48 phoneme sets modelled by HMMs of 3 non-skipping states and 3 mixture Gaussian. The phoneme recognition accuracy under no grammar condition is improved by .9%, as shown in table 2.

	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.

To measure the effectiveness of prosody dependent acoustic modelling, we conducted phoneme recognition experiments on Radio New Corpus with no grammar used. Table 3 shows the percent phoneme recognition accuracy (PRA) for various types of boundary phone models as defined in Table 1. Boundary phones and non-boundary phones are counted as different symbols in this result. Note that the figures row-wise are not comparable because they are measured under different phoneme sets of different sizes. The figures in column IND are the percent PRAs achieved with IND phoneme set in which boundary and non-boundary phones are different logically but are the same physically; while the numbers in column PD are the percent PRAs achieved with the prosody dependent phonemes having untied duration PMFs. Note that the boundary phones and non-boundary phones only differ in their duration PMFs in our study, hence the total number of parameters of these prosody dependent models is not significantly increased over that of the baseline IND models. Compare column-wise, we see that the PRAs are drastically improved in FV, FC and FVFC for both HMMs and EDHMMs. This indicates that the lengthening in phrase final rhymes can be reliably modelled by HMM. Conversely, PRAs degrade in all the sets that contains phrase initial phones. This indicates that there are no systematic duration variation in the phrase initial positions that can be reliably modelled by HMM. It can be concluded from these results

	HMM		EDHMM	
	Corr.	Acc.	Corr.	Acc.
IND	52.60	36.77	52.59	37.15
FV	52.64	36.81	52.58	37.25
FC	52.62	36.73	52.74	37.10
FVFC	52.67	36.76	52.79	37.19
IC	50.55	34.64	50.79	34.24
IV	50.6	34.68	50.69	35.26
IVIC	48.59	32.69	48.89	33.50
ICFV	52.65	36.63	52.79	37.18
IPFP	52.61	36.43	52.80	37.05

Table 4: Phoneme Recognition with boundary and non-boundary distinction ignored.

AM	LM	HMM		EDHMM	
		Corr.	Acc.	Corr.	Acc.
IND	IND	76.17	74.89	76.99	75.15
IND	PF	76.75	76.36	77.56	75.60
IND	PI	77.36	75.40	77.52	75.40
IND	PP	77.40	75.60	77.56	77.68
FVFC	IND	76.17	75.03	76.58	75.23
FVFC	PF	76.83	75.44	77.44	75.52
FVFC	PI	77.36	75.32	77.44	75.44
FVFC	PP	76.95	75.60	77.28	75.85

Table 5: Word Recognition using IND and FVFC models in combining with IND, PF, PI and PP language models.

that duration is very important for prosody dependent phoneme recognition. Table 4 shows the net phoneme recognition results in which all B4 prefixes and postfixes are ignored in counting the results. Now, the number row-wise is comparable and we see that phoneme recognition accuracy has been improved in all phrase final type of phoneme modelling with maximum improvement .5% appearing in FV models.

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 four 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 four types of language models are denoted as IND, PF, PI and PP. Here, IND denotes a LM that is completely prosody independent; PF denotes a LM that distinctively models phrase final words; PI denotes a LM that distinctively models the phrase initial words; and PP is the LM that has the maximal prosody dependence in which all 3 types of words: phrase medial, phrase initial and phrase final are distinguished. As can be seen in Table 5, the word recognition accuracy (WRA) of FVFC+PP+EDHMM has improved about 1% over the baseline system IND+IND+HMM. The improvement bought by acoustic modelling is not very evident in this results because the LM models has dominate effectiveness on this database due to the repetitive sentences in training and testing set. We would expect the effectiveness of acoustic modelling to be more evident on larger unbiased database.

Table 6 and Table 7 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. Intonational phrase boundary recognition is potentially a difficult task because only less than 1/5 of the word boundaries are intonational phrase boundaries. Simply setting all word boundaries to be B0 will give an accuracy of over 80%. Nevertheless, we achieved over 1% improvement in both types of boundary recognition in FVFC+PP+EDHMM over the IND+IND+HMM.

7. 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 radio news, the duration lengthening in phrase final syllable rhymes can be utilized to improve acoustic modelling. The prosody dependent speech recognition we proposed improves both word recognition accuracy and boundary recognition accuracy by 1%.

AM	LM	HMM		EDHMM	
		Corr.	Acc.	Corr.	Acc.
IND	IND	84.55	84.47	84.72	84.43
IND	PF	84.63	84.43	84.63	84.43
IND	PI	87.94	85.33	88.07	85.37
IND	PP	88.07	85.33	88.07	85.37
FVFC	IND	84.59	84.43	84.72	84.43
FVFC	PF	84.63	84.43	84.63	84.47
FVFC	PI	87.82	85.37	87.90	85.41
FVFC	PP	88.11	85.49	88.25	85.49

Table 6: Phrase Initial Boundary Recognition.

AM	LM	HMM		EDHMM	
		Corr.	Acc.	Corr.	Acc.
IND	IND	84.55	84.47	84.72	84.43
IND	PF	84.63	84.43	84.63	84.43
IND	PI	84.68	84.55	84.76	84.47
IND	PP	87.86	85.33	88.97	85.53
FVFC	IND	84.59	84.43	84.59	84.47
FVFC	PF	84.63	84.43	84.63	84.47
FVFC	PI	84.68	84.55	84.76	84.47
FVFC	PP	88.15	85.49	88.48	85.62

Table 7: Phrase Final boundary Recognition.

8. References

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

[2] 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.

[3] 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.

[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] 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.

[6] 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.

[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] 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.