Proceedings of HLT-NAACL 2003
Student Research Workshop , pp. 7-12
Edmonton, May-June 2003

The Importance of Prosodic Factors in Phoneme Modeling with Applications to Speech Recognition

Sarah Borys

Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign, Urbana, IL 61901

Abstract

This paper tests speech recognition using prosody dependent allophone models. The log likehoods of various prosodically labeled phonemes are calculated using Baum-Welsh re-estimation. These log likehoods are then compared to log likehoods of non-prosodically labeled phonemes. Based on the comparison of these log likehoods, it can be concluded that modeling all prosodic information directly in the vowel model leads to improvement in the model. Consonants, on the other hand, split naturally into three categories, strengthened, lengthened and neutral.

1. Introduction

Prosody is an important factor in how humans interpret speech. The same word string can have different meanings depending on the way it is said. Many linguists have performed extensive studies of prosody and of the effects of prosodic factors on spoken language.

In his dissertation, Cho (2001) investigates how phonetic features are conditioned by prosodic factors by examining pre-boundary, post-boundary, and accented syllables. Cho reports that boundary induced articulatory strengthening occurs in phrase final vowel positions and phrase initial consonant positions. Phrase initial vowels are also more susceptible to coarticulation than phrase final vowels. Cho also hypothesizes that accented syllables are characterized primarily by sonority expansion. An accented vowel is usually not affected by coarticulation with a neighboring vowel. Strengthening effects caused by boundaries and accents cannot be considered the same and Cho discusses several differences between boundary and accent strengthening effects.

In a study performed by Edwards et al (1991), the effect of final lengthening at prosodic boundaries was examined by studying articulator movement patterns. It was found that decreasing intragestural stiffness slows down the syllable, affecting the tempo of the spoken word, causing the syllable to be lengthened. The changing of intergestural phrasing also affects the syllable duration by decreasing the overlap of a vowel gesture with a consonant gesture. This increases the duration of accented syllables comparatively to unaccented syllables and causes the accented syllable to be strengthened.

De Jong (1994) investigated the supraglottal correlates of linguistic prominence in English. De Jong suggests that stress involves a localized shift toward hyperarticulated speech. An increase in the duration in the closure and in the aspiration of initial voiceless stops was observed along with an increase in duration of prevoicing in initial voiced stops in stressed syllables.

Fougeron and Keating (1997) report that on the edges of prosodic phrase boundaries, final vowels and initial consonants have less reduced lingual articulation. The differences in articulation were manifested in the linguopalatal contact of boundary consonants and vowels. The linguopalatal contact of both consonants and vowels relates directly to the type and size of phrase boundary. Boundary type and size also appear to effect the acoustic duration of post-boundary consonants.

Wightman et al (1992) report that there is segmental lengthening in the rhyme of a syllable that directly precedes a phrase boundary. Wightman examines the effect of duration and pause on boundary words and shows that speaking rate effects the distribution of phoneme duration. The lengthening effects of pre-boundary syllables can be used to distinguish several different types of phrase boundaries.

These results show that prosody can cause variations not just in pitch, but also in the articulation of phonetic contrasts in different phonemes. These variations can be modeled as a part of the phoneme definition in an automatic speech recognition (ASR) system. However, the question is whether or not modeling prosodic factors with phonemes would lead to improvements in the quality of the phoneme model and thus lead to improvements in both the correctness and accuracy in an ASR system.

Most modern speech recognizers function by breaking words up into mathematical features. The recognizer then determines the most likely occurring set

Consonants			Vowels	
b	ch	d	aa	ae
dh	f	g	ah	ao
hh	jh	k	aw	ax
l	m	n	ay	eh
p	r	s	el	er
sh	t	v	ey	ih
w	y	z	iy	ow
			oy	uh
			uw	

Figure 1. This figure contains a chart of the 38 different non-prosodically distinguished phonemes used for experimentation.

phn : phrase medial
phn! : phrase medial, accented
phnB4 : phrase final, unaccented
phnB4! : phrase final, accented
B4phn : phrase initial, unaccented
B4phn! : phrase initial, accented

Figure 2. The different prosodic labels. "Phn" represents some generic phoneme.

of phonemes by comparing these extracted features with its own phoneme models. Phonemes are usually modeled using hidden Markov Models (HMMs). Once the recognizer has identified a set of the most likely occurring phonemes, it then uses a dictionary to match a word or group of words to that set.

Prosody can be incorporated into the phoneme model by allowing two different HMMs to represent a single phoneme. One HMM would need to represent the prosody independent version of the phoneme while the other would represent the phoneme in some prosodic context. This could allow the recognizer to do things such as distinguish between accented and unaccented phonemes or distinguish between boundary and non-boundary phonemes. Allowing the recognizer to make such a distinction may reduce the confusability of certain phoneme groups, which in turn could allow for increased recognition rates.

The goal of this research is to not only determine if the inclusion of prosody in the phoneme model causes improvement in the model, but also to determine which prosodic factors to model and the best way to model them. This will be accomplished by first splitting phonemes into different prosodically varying groups and then by comparing the log probability of the occurrence of each phoneme in those different groups. Because prosody causes noticeable variations in speech, a phoneme model that includes prosodic factors should differ from models of the same phoneme that do not. This difference will prove to be significant enough to show that prosodic factors should be taken into account for a more accurate phoneme model.

2. The Database

Boston University's Radio News Corpus (1995) was used for all experiments. The speakers from this corpus that were analyzed were F1A, F2B, and M2B. The usable data from these three speakers consisted of 259 wav files containing 18270 words. All the wav files that were used were accompanied by two types of prosodic transcription files, .brk and .ton files.

The corpus was labeled according to the ToBI standard. Silverman et al (1992) explain the labeling system in detail. It will not be described in this paper.

The .brk files specify a ToBI break index (0-4) for every spoken word in the associated wav file. For the experiments, the only boundary distinguished was the intonational phrase boundary (ToBI index 4). All other boundary types (indices 0-3) were grouped together. There were 3855 intonational phrase boundaries in the data set.

The .ton files label the times in which an accented vowel occurs. The most abundant accent label was H* which occurs in a ratio of about 10 H* for every single L*. Other accent types do occur, but most include H* in bitonal accent.

3. Prosodic Annotation

The set of 38 different phonemes, shown in figure 1, were used in the experiments.

3.1 Allophone Modeling

Recognition experiments were preformed for four different allophone sets:

- Tied
- Accent
- Boundary
- Untied

The Tied set contained no prosodically labeled data.

The Accent set contained monophones that were split into two groups, accented and unaccented. Phonemes were not distinguished on the basis of phrase position.

	Tied	Accent	Boundary	Untied
Monophone Group	All Cons.	All Cons.	All Cons.	All Cons.
	After Vowel	After Vowel	After Vowel	After Vowel
	Before Vowel	Before Vowel	Before Vowel	Before Vowel
	Vowels	Vowels	Vowels	Vowels

Figure 3. The sixteen experimental conditions

The Boundary set modeled monophones as phrase initial, phrase medial, or phrase final. Accented phonemes were not distinguished from unaccented phonemes.

The Untied set distinguish phonemes by both phrasal position and accentuation. A monophone in this group could be labeled as phrase medial, phrase medial accented, phrase initial, phrase initial accented, phrase final or phrase final accented.

3.2 Allophone Definitions

Figure 2 contains the six different labels used to represent the allophones of a single imaginary phoneme "phn."

A phrase final phoneme was considered to be any phoneme that occurred in the nucleus or coda of the final syllable of a word directly preceding an intonational phrase boundary. Phrase initial phonemes, on the other hand, were considered to be any phoneme in the onset or nucleus of the initial syllable of a word that followed an intonational phrase boundary. Phase medial phonemes were considered to be any other phoneme.

An accented vowel was the lexically stressed vowel in a word containing a transcribed pitch accent. Because accented consonants are not clearly defined, three different labeled sets of accented consonants were developed:

- All Consonants
- After Vowel
- Before Vowel

All Consonants considered every consonant in a syllable with an accented vowel to also be accented. After Vowel considered as accented only the coda consonants. Before Vowel recognized only the onset consonants of the accented syllable as being accented. Accents were considered to be limited to a single syllable.

Because there were three different groups of accented consonants and because there is only one way a vowel can be labeled as accented, vowels were separated into a fourth group of their own, entitled Vowels. The four groups along with the four different allophone models lead to the sixteen experimental conditions illustrated in figure 3.

beyond b iy y aa n d
beyond! b iy y aa! n! d!
beyondB4 b iy y aaB4 nB4 dB4
beyondB4! b iy y aaB4! nB4! dB4!
B4beyond B4b B4iy y aa n d
B4beyond! B4b B4iy y aa! n! d!

Figure 4. An example of each of the six word types defined with Untied allophones for the After Vowel experimental condition. Boundary allophones could only be used to define three distinct word types, Accent only two, and Tied only one.

a.
0 370000 B4in
370000 760000 nineteen!
760000 1150000 seventy
1150000 1680000 sixB4
1680000 2310000 B4democratic!
2310000 2680000 governor

b.
600000 1600000 w
1600000 2400000 aa!
2400000 2900000 n!
2900000 3800000 t
3800000 4900000 axB4
4900000 5300000 dB4

Figure 5a. An example Untied word level transcription
b. An example Untied phone level transcription for the After Vowel accent condition. The transcribed word is "wanted."

3.3 Dictionaries and Transcription Types

Each experimental condition required its own dictionary and transcription. Just as each phoneme had six distinct allophones, each word had six distinct types. A word could be phrase initial, medial or final and accented or unaccented. Each word type had its own definition. An example dictionary is shown in figure 4.

Every experimental condition had both a word level transcription and a phone level transcription. Figure 5 shows an example of the two different levels of transcription files.

4. Experiments

All Consonants		After Vowel		Before Vowel		Vowels	
Merge	Separate	Merge	Separate	Merge	Separate	Merge	Separate
ch	b	dhB4	b	B4d	b	aoB4	aa
dB4,	B4b	gB4	ch	B4f	B4b	ax	aaB4
B4d	d	jhB4	d	B4g	ch	B4eh	B4aa
dhB4	dh	kB4	dB4	B4k	d	B4ey	ae
B4dh	f	lB4	dh	B4m	dh	B4ow	aeB4
fB4	g	mB4	f	B4n	f	uh	B4ae
B4f	hh	nB4	fB4	B4p	g	uhB4	ah
gB4	jh	pB4	g	B4s	hh	B4uh	ahB4
B4g	jhB4	pB4	jh	B4w	B4hh	B4uw	ao
jhB4	k	sB4	k	z	jh		aoB4
kB4	l	sh	l		k		aw
B4k	m	tB4	m		l		ay
lB4	n	v	n		m		ayB4
mB4	nB4		p		n		eh
B4m	p		r		p		ehB4
pB4	r		rB4		r		ey
B4p	rB4		s		B4r		eyB4
sB4	B4r		sh		s		ih
B4s	s		t		sh		ihB4
tB4	sh		vB4		t		B4ih
v	t		y		B4t		iy
B4w	tB4		z		v		iyB4
z	vB4		zB4		w		B4iy
	B4v				y		ow
	w						owB4
	y						oy
	zB4						uw
							uwB4

Table 1. The results of experiments for the Accented allophone sets. The "Merge" column lists phonemes with WA ≥ LL. The "Separate" column indicates phonemes where WA < LL. Due to the relatively small size of the data set, several phonemes are missing from the table.

Experiments were performed using the Hidden Markov Toolkit (HTK), which is distributed by the University of Cambridge (2002). Phonemes were modeled using a three-state HMM with no emitting start and end states. Each emitting state consisted of three mixture Gaussians and no state skipping was allowed.

4.1 Experimental Procedure

The Radio News Corpus data was divided into 2 sets: a training set and a test set. The test set was approximately 10% of the size of the training set. The experimental procedure was completed for sixteen experimental conditions.

The experimental procedure can be divided into two steps. In step one, the training data was used to re-estimate the HMM definitions for each phoneme. Re-estimation was performed with the HTK tool HRest, which uses Baum-Welsh re-estimation described in detail in the HTK book available from Cambridge University (2002). HMM parameters were re-estimated until either the log likehood converged or HRest had performed 100 iterations of the re-estimation algorithm.

In the second step of the experiments, HRest was used to perform a single iteration of the re-estimation algorithm on the test data using the HMM definitions that were updated from the re-estimation of the training set. During re-estimation, the log likehoods of each phoneme were output and saved for later comparisons.

4.2 Post Processing

Once all the log likehoods had been recorded, the Untied allophone sets were used as a basis to determine if the considered monophones were better modeled as prosody independent or prosody dependent. To determine the best modeling strategy for a particular monophone, six different weighted averages (WA's) were calculated from the Untied log likehoods and compared to the computed log likehoods of the Boundary, Accent and Tied models.

a.

	Initial	Medial	Final
Accented	1		3
Unaccented		2	

b.

	Initial	Medial	Final
Accented	1	2	3
Unaccented	4	5	6

Figure 6a. The proposed modeling of consonants. 1 = Strengthened, 2 = Neutral, 3 = Lengthened
b. The proposed modeling of Vowels. Numbers 1-6 indicate six different distinguishable prosodic types

The following three formulas were used to calculate the WA's of the Untied set for comparison with the Boundary set computed value:

$$WA_{PM} = L_{phn}W_{phn} + L_{\ phn!}W_{phn!}$$

$$WA_{PI} = L_{B4phn}W_{B4phn} + L_{B4phn!}W_{B4phn!}$$

$$WA_{PF} = L_{\ phnB4}W_{phnB4} + L_{phnB4!}W_{phnB4!}$$

where PM, PI, and PF stand for phrase medial, initial and final, respectively. L_x represents the computed log likehood of the allophone label x in the Untied allophone set, and W_x represents the frequency of that x.

W_x, where x is representative of any of the six types of prosodically labeled monophones, is computed by the following formula:

$$W_x = num_x / TOTAL$$

where num_x represents the number of examples of the token x, and TOTAL is the sum of all the different phoneme tokens being taken into account for the computation of WA of some set of phonemes.

The two formulas used in calculating the WA's for comparison with the Accent allophone set are as follows:

$$WA_U = L_{phn}W_{phn} + L_{B4phn}W_{B4phn} + L_{phnB4}W_{phnB4}$$

$$WA_A = L_{phn!}W_{phn!} + L_{B4phn!}W_{B4phn!} + L_{phnB4!}W_{phnB4!}$$

where WA_U and WA_A are the weighted averages of log likehoods for the accented and unaccented tokens respectively.

The WA compared to the Tied set was computed as follows:

$$WA_T = L_{phn!}W_{phn!} + L_{B4phn!}W_{B4phn!} + L_{phnB4!}W_{phnB4!} + L_{phn}W_{phn} + L_{B4phn}W_{B4phn} + L_{phnB4}W_{phnB4}$$

where WA_T is the weighted average of all of the phonemes in the Untied model.

The weighted averages were then compared to the log likehoods using the following algorithm:

if (WA < LL), then split using prosodic labels

if (WA ≥ LL), then do not split using prosodic labels

LL is the log likehood computed using HRest.

5. Results

For each prosodic variable (phrasal position or accent), tables were constructed listing the preferred tying of phonemes based on the log likehood results. Table 1, for example, lists all phonemes that should be tied on the basis of accent and those that should not. Similar tables exist for phrasal position and for the combination of both accent and phrasal position. Examples of certain phonemes are not present due to the relatively small size of the data set.

Experimental results varied greatly between consonants and vowels. For consonants, there appeared to be an improvement in the model when phonemes are distinguished by phrasal position. Separation of accented and unaccented phrase initial consonants yielded no improvement to the model for most consonants. This implies that phrase initial accented and phrase initial unaccented phonemes should be merged into a single token. Accented consonants are also not benefited by positional information. Results indicate that phrase initial, medial and final accented phonemes can be merged together. Figure 6a illustrates a proposed model for the prosodic labeling of consonants based on these results.

For vowels, a model showed improvement when the phoneme was separated into phrase initial, medial and final tokens. Vowel phoneme models also showed improvement when separated by accent. The accent on a vowel appears to be important regardless of phrasal position. These results suggest a six-way distinction should be used when modeling vowels and the proposed model is illustrated in figure 6b.

6. Conclusion

While the data used for these experiments was sparse for certain phonemes, many of the phoneme models tested showed improvement when prosody was incorporated directly into the HMM definition. Analysis of experimental results led to two different

proposals for the modeling of consonants and vowels. Verifying that the proposed models are indeed an improvement over standard phoneme modeling will be a goal of future work.

Acknowledgements

This work could not have been completed without the help and guidance of Professor Mark Hasegawa-Johnson and Professor Jennifer Cole.

7. References

Cho, T. 2001 *Effects of Prosody on Articulation in English.* Ph.D. dissertation, UCLA.

De Jong, Kenneth (1995) "The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation," *JASA,* vol.97(1), pp. 491-504.

Edwards, Jan. Beckman, Mary, & Fletcher, Janet. 1991 "The articulatory kinematics of final lengthening," *JASA* 89(1), pp. 369-382.

Fougeron, P. & Keating, P. 1997 "Articulatory strengthening at edges of prosodic domains," *JASA* 101(6), pp. 3728-3740.

Ostendorf, M., Price, P.J., Shattuck-Hufnagel, S. 1995. "The Boston University Radio News Corpus," Boston University Technical Report No ECS-95-001, <http://ssli.ee.Washington.edu/papers/radionews-tech.ps>.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M. Wighnman, C. Price, P., Pierrehumbert, J., Hirschberg, J., 1992, "ToBI, a standard for labeling English" *ICSLP*, vol. 2, pp867-870

The University of Cambridge Engineering Department, 2002. "http://htk.eng.cam.ac.uk/".

Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., & Price, P. J. 1992. "Segmental durations in the vicinity of prosodic phrase boundaries," *J. Acoust. Soc. Am.*, vol. 91, no. 3, pp 1707-17