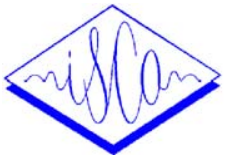

INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

MODELING AND RECOGNITION OF PHONETIC AND PROSODIC FACTORS FOR IMPROVEMENTS TO ACOUSTIC SPEECH RECOGNITION MODELS

Sarah Borys[1], *Mark Hasegawa-Johnson*[1], *Jennifer Cole*[2], *Aaron Cohen*[1]

1. Department of Electrical and Computer Engineering
2. Department of Linguistics
University of Illinois at Urbana-Champaign, Urbana, IL 61801
{sborys, jhasegaw, jscole, ascohen}@uiuc.edu

10.21437/Interspeech.2004-756

Abstract

This paper examines the usefulness of including prosodic and phonetic context information in the phoneme model of a speech recognizer. This is done by creating a series of prosodic and phonetic models and then comparing the mutual information between the observations and each possible context variable. Prosodic variables show improvement less often than phone context variables, however, prosodic variables generally show a larger increase in mutual information. A recognizer with allophones defined using the maximum mutual information prosodic and phonetic variables outperforms a recognizer with allophones defined exclusively using phonetic variables.

1. Introduction

Prosody provides important cues to humans for speech understanding because not only does it change word meaning, but it also affects the quality of phones. If computers were provided knowledge of prosody, would they be able to better recognize speech?

Many linguists have examined the effects of prosody on speech. Wightman et al. [1] showed that phrase boundary depth affects the distribution of phoneme duration in the preceding syllable rhyme. Fougeron and Keating [2] have shown that on the edges of prosodic phrase boundaries, final vowels and initial consonants have less reduced lingual articulation. De Jong [3] observed an increase in duration of prevoicing in initial voiced stops in stressed syllables. Edwards et al. [4] found that changing intergestural phrasing reduces the overlap of a vowel gesture with a consonant gesture, causing an increase in duration and a strengthing effect for accented syllables. Cho [5] has shown that accented vowels are not usually affected by coarticulation with neighboring vowels and phrase initial vowels are susceptable to coarticulation. Cho notes that boundary induced articulatory strengthening occurs in phrase final vowel positions and phrase initial consonant positions.

Researchers have also shown that machines are able to detect the changes in speech induced by prosody. Wightman and Ostendorf [6] present two algorithms that can detect and label prosodic phrase boundaries with over 90% accuracy. In [7] Wightman and Ostendorf use a modification of the algorithms used in [6] to detect prominences with accuracy of 86% and boundary tones with an accuracy of 77%.

The goal of this paper is to examine two different prosodic factors (intonational phrase structure and pitch accent), one syntactic factor (function vs. content words), and ten phone contexts to determine which of these factors are most useful to computers for the purpose of speech recognition. Criteria for usefulness will include mutual information between context variable and acoustic observations, and word recognition accuracy of a speech recognizer built with appropriate context-dependent allophones.

	Stop	Fricative	Liquid	Nasal	Vowel
Left	LS	LF	LL	LN	LV
Right	RS	RF	RL	RN	RV

Table 1: The ten non-prosodic, phone context dependent allophone sets and their abbreviations. The words "Left" and "Right" refer to which neighboring phone is being considered. The phonetic classes, stop, fricative, liquid, nasal or vowel, specify the type of the considered neighboring phone.

Prosody Independent	IND
Accent Onset	Onset
Accent Coda	Coda
Accent All	All
Content-Function	CF
Phrase Final	PF
Phrase Initial	PI

Table 2: The prosodic allophone sets.

2. Phone Splitting

Determining whether or not prosodic and phone context (PC) factors would be useful to the recognizer was accomplished by examining the log likelihoods of prosodic and PC independent phones and comparing them to log likelihoods of phone models that have been split via a binary prosodic or phonetic distinction. Binary distinctions are listed in Tables 1 and 2.

The monophone set used for phone splitting and recognition experiments was a version of the Sphinx monophone set [8]. The base set of monophones contained 46 distinct phones and silence. This base set is referred to as the independent (IND) set.

The AC set split phones into two groups, accented and unaccented. An accented vowel was defined to be the vowel in the lexically stressed syllable of a word with a transcribed pitch

accent. The phonetics literature suggests that consonants in the onset of an accented syllable are more clearly enunciated than consonants in the onsets of other syllables (e.g., DeJong, [3], Cole et al. [9]), but does not specify whether or not rhyme consonants are similarly hyper-articulated, so therefore, in order to determine the full effect of accentuation on consonants, three sub-sets of accented consonants were created. These subsets were Coda, Onset and All. The Coda subset defined an accented consonant to be any consonant occurring in the coda of the syllable containing an accented vowel. The Onset subset defined an accented consonant to be any consonant contained within the onset of the syllable containing the accented vowel. The All subset combined both Onset and Coda.

The CF allophone set distinguished phones as being either a content phone or a function phone. A function phone is a phone that occurs in a function word.

The PF allophone set split phones into "phrase final" and "non-final" phones. A phone was considered to be phrase final if it occurred in the nucleus or coda of the final syllable in a word that preceded an intonational phrase boundary.

The PI allophone set separated phones into the groups "phrase initial" and "non-initial." A phrase initial phone could occur only in the onset or nucleus of the first syllable in a word that followed directly after an intonational phrase boundary.

3. Experiments

The Boston University Radio News Corpus [10] was used for experimentation. Data was used from speakers F1A, M1B, F2B, M2B, and F3A.

The dataset contained over three hours of data divided between a training set and a test set. The test set was approximately 10% the size of the training set. The training set included 23,103 words in 272 files. 11,386 of the words were accented. The training set also included 3829 intonational phrase boundaries. The database was judged to be too small to train a speaker-independent recognizer (removing two talkers for testing purposes would have left too little data in the training database), therefore all talkers were represented in both training and test databases.

3.1. Log Likelihood Comparison

The goal of this experiment was to determine which prosodic and phonetic factors could be used to improve speech recognition results. This was accomplished through comparing the log likelihoods of the IND allophones to the log likelihoods of the corresponding split allophones for each predefined allophone set shown in Tables 1 and 2.

The log likelihoods were found using HTK, the Hidden Markov Toolkit [11].

Once all models had been trained, the HRest function was used to excise examples of each allophone from the test corpus, and to compute the total log probability of the test examples given the HMM parameters. Consider phoneme q_0, represented by N_0 different examples in the test corpus, $X_0 = \{x_1, \ldots, x_{N_0}\}$, where x_k is a matrix containing the sequence of Mel frequency cepstral coefficient (MFCC) vectors spanning the entire duration of the kth test-corpus example of phoneme q_0. The standard definition of language model perplexity is based on a measure of cross-entropy, $\hat{\mathcal{H}}(x|q = q_0)$, and can be approximated as

$$\hat{\mathcal{H}}(x|q = q_0) = -\frac{1}{N_0}\sum_{k=1}^{N_0} \log_2 \hat{p}(x_k|\Lambda_0) \quad (1)$$

where $\hat{p}(x_k|\Lambda_0)$ is the likelihood extimated by an HMM with parameters Λ_0.

Similarly, the cross-conditional differential entropy of x given $q = q_0$, and given knowledge of the binary context variable $v \in \{-1, 1\}$, is defined to be

$$\hat{\mathcal{H}}(x|q = q_0, v) = -\frac{1}{N_0}\left(\sum_{k=1}^{N_1} \log \hat{p}(x_k|\Lambda_{0,-1}) + \sum_{k=N_1+1}^{N_0} \log \hat{p}(x_k|\Lambda_{0,1})\right) \quad (2)$$

where $\Lambda_{0,-1}$ are parameters of the HMM trained to represent allophones of q_0 in the context $v = -1$, $\Lambda_{0,1}$ represent q_0 in context $v = 1$, and we assume that the test database has been sorted so that tokens $\{x_1, \ldots, x_{N_1}\}$ are in context $v = -1$ and all others are in context $v = 1$.

Given Eqs. 1 and 2, the mutual information between x and v, conditioned on phoneme q_0, is defined in the usual way:

$$\hat{\mathcal{I}}(x, v|q = q_0) = \hat{\mathcal{H}}(x|q = q_0) - \hat{\mathcal{H}}(x|q = q_0, v) \quad (3)$$

Eq. 1 can be computed by using the HRest program to train Λ_0 on training data, then running HRest once on the test corpus, and discarding the re-estimated parameters; as an auxiliary output, HRest computes the average log-likelihood of the test tokens given Λ_0, equivalent to $\hat{\mathcal{H}}(x|q = q_0)$. Eq. 2 may be computed by using HRest to train and test two context-dependent allophone models, and computing their weighted average.

3.2. Speech Recognition

The IND monophone set contained only 47 monophones, including silence. The result of splitting any of these monophones into prosodic allophones is that the number of phones, and thus the number of model parameters, will increase. An increase in model parameters will have a tendency to favor prosodic splitting, so therefore, in order to achieve an accurate comparison between a prosody-independent and a prosody-dependent recognizer, the number of phones in the prosody-independent recognizer should be increased to match the number in the prosody dependent recognizer.

In our first attempts to build prosody-dependent and prosody-independent recognizers, we tried to use the HTK tree-based clustering utilities to cluster individual states in each allophone HMM. Tree-based clustering algorithms failed to converge using the prosodically transcribed subset of Radio News, apparently because of insufficient data (for example, the Baum-Welch algorithm would frequently allocate fewer than two frames of training data to some triphone state, resulting in re-estimation failure). For this reason we developed a top-down algorithm for clustering allophone models, as follows.

The mutual information measure in Eq. 3 was computed for all of the phonetic context variables listed in Table 1. For each phoneme, the two context variables with the highest $\mathcal{I}(x, v|q = q_0)$ were selected, defining three prosody-independent allophones of phoneme q_0 ($v_1 = 1$, $(v_1, v_2) = (1, -1)$, $(v_1, v_2) = (-1, -1)$). The prosody-independent speech recognizer was

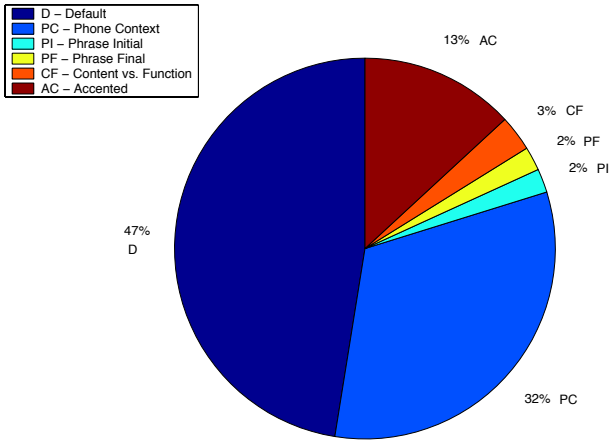

Figure 1: Results of retranscribing the training data using the prosody-dependent allophone set. See text.

trained using 139 prosody-independent allophone models (three allophones of each of the 46 IND monophones and silence). The prosody-dependent recognizer was then built by keeping the number of model parameters constant and replacing the PC dependent model for a given phone with a prosody dependent model provided that the replacement model showed an improved log likelihood. If a phone had no prosodic splits that allowed for improvement in the log likelihood of the model, then the PC splits were not replaced.

Figure 1 shows what percentage of the 82,729 monophone training set came from each split for the prosody-dependent recognizer. In the figure, PC phones are phones for which there was not significant improvements in the likelihood function due to prosody, but there was still improvement due to a given phone context. Default phones are phones that did not show improvements for either the PC or prosody questions. The prosody-dependent phones are sorted by question, PI, PF, CF and AC, where AC combined the three splits Onset, Coda and All. 20% of the segments were positive examples of a prosodic-context dependent allophone judged to be important by the splitting algorithm (PI, PF, CF, AC). 32% of the segments were positive examples of a phone-context dependent allophone. 47% used the default context-independent allophone label.

4. Results

4.1. Log Likelihood Comparison

In general, splitting monophones based on prosodic or phone context improves the phoneme model. Phone context splits seem to affect all linguistic phonetic groups equally, while prosodic splits tend to favor different types of phones. Vowels showed improvement over almost every defined prosodic category. For consonants, the effects of prosodic splitting are most significant for plosives and nasals. Prosody appears to have no or very little effect on fricatives and glides. Table 3 sumarizes these results.

As can be seen from Table 3, there are fewer prosodic splits that allow for model improvement than there are phone context splits. When there is an improvement due to prosodic splitting, the mutual information resulting from these splits suggest that the inclusion of prosodic factors in the model will allow for greater model improvement than will the inclusion of phonetic context factors. Figure 2 gives statistics for each linguistic phonetic category.

	% Prosodic	% Phone
Vowels	73.2	97.5
Stops	58.5	91.7
Fricatives	27.3	87.8
Nasals	76.5	94.9
Glides	8.3	92.0

Table 3: Percentages of prosodic or phone context splits that showed improvement for each linguistic phonetic group.

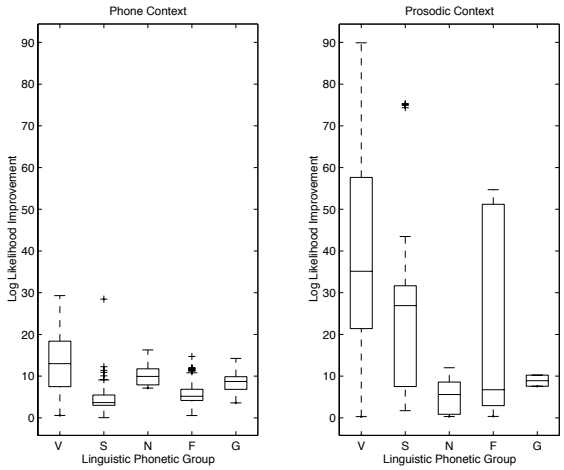

Figure 2: Box and whisker plot of likelihood function improvements for prosodic and PC splits for each linguistic phonetic category.

4.2. Fixed Parameter Recognizer

Word recognition accuracy of the prosody-dependent and prosody-independent recognizers are shown in Table 4. As seen in the table, when prosodic factors are used as model parameters, the number of words correctly identified by the recognizer increases by 9.85%. The accuracy of the prosody based word recognizer is greater than that of the prosody independent recognizer by 10.83%.

The main effect of incorporating prosody into the recognizer is that the number of word substitution errors is drastically reduced. Substitutions are reduced by 35% (relative) between the prosody independent and prosody dependent recognizers. Word insertions and deletions made by the prosody dependent recognizer are also reduced by 27% and 15% respectively from the number of insertions and deletions made by the independent recognizer.

	% Correct	Accuracy
Independent	67.31	63.75
Dependent	77.16	74.18

Table 4: Word recognition accuracy for the prosody independent and prosody dependent recognizers.

5. Conclusion

Both prosodic and phone context information can be included in the phoneme model. When included, this information causes the log likelihood of the model to improve. Prosodic information is more useful to the model than is phonetic context information for most phonemes. This has been shown both in experiments that compare the log likelihoods of phone context-dependent models to those of prosodic context-dependent models and also in experiments that included both prosodic and phonetic context models in two different speech recognizers.

Prosody is essential for human understanding and interpretation of speech. This work has shown that prosody is also significantly important to computer recognition and understanding of speech.

6. Acknowledgements

This work was supported in part by NSF award number 0132900, and in part by a grant from the University of Illinois Critical Research Initiative. Statements in this paper reflect the opinions and conclusions of the authors, and are not endorsed by the NSF or the University of Illinois.

7. References

[1] Wightman, C. W., Shattuck-Hufnagel, S. Ostendorf, M., Price, P. J. *Segmental durations in the vicinity of prosodic phrase boundaries*. Journal of the Acoustical Society of America, 1992. vol. 91(3), pp: 1707-17.

[2] Fougeron, P., Keating, P. *Articulatory strengthening at the edges of prosodic domains*. Journal of the Acoustical Society of America, 1997. vol 101(6), pp: 3728-3740.

[3] De Jong, K., *The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation*. Journal of the Acoustical Society of America, 1995. vol.97(1), pp: 491-504.

[4] Edwards, J., Beckman, M., Fletcher, J. *The articulatory kinematics of final lengthening*. Journal of the Acoustical Society of America, 1991. vol 89(1), pp: 369-382.

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

[6] C. W. Wightman, M. Ostendorf, *Automatic recognition of prosodic phrases*. International Conference on Acoustics, speech, and signal Processing, April 1991, pp: 321-324.

[7] C. W. Wightman, M. Ostendorf, *Automatic recognition of intonational features*. IEEE International Conference on Acoustics, Speech, and Signal Processing, March 1992, pp: 221-224.

[8] Lee, K., Hon, H., Reddy, R. *An overview of the SPHINX speech recognition system*. IEEE Transactions on Acoustics, Speech, and Signal Processing, 38(1). January 1990.

[9] Cole, J., Choi, H., Kim, H. and Hasegawa-Johnson, M. *The effect of accent on the acoustic cues to stop voicing in Radio News speech*. International Congress of Phonetic Sciences, 2003.

[10] Ostendorf, M., Price, P. J., Shattuck-Hufnagel, S. *The Boston University radio news corpus*. Linguistic Data Consortium, Philadelphia, PA. 1995.

[11] Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Valtchev, V., Woodland P. *The HTK book*. Cambridge University Engineering Department, http://htk.eng.cam.ac.uk/, 2002.