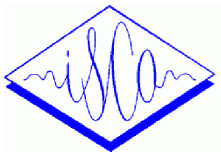

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Using Prominence and Phrasing Predictions to Improve Weighted Dictionary Pronunciation Models

Andrew Rosenberg[1]

[1]Department of Computer Science, Queens College (CUNY), Flushing, NY, USA
andrew@cs.qc.cuny.edu

Abstract

Prosody impacts the pronunciation variation of lexical items in a number of ways. Accented syllables tend to be pronounced with their canonical (dictionary) vowel. Deaccented vowels are more likely to be reduced. Coarticulatory influences rarely span intonational phrase boundaries. In this work, we investigate the use of automatically generated prosodic hypotheses to improve a weighted dictionary pronunciation model. We use the phonemically transcribed, Buckeye Corpus for this investigation. We find that predictions of pitch accent and intonational phrase boundaries can be used to lower pronunciation model perplexity.

Index Terms: Pronunciation Modeling, Prosody, Prominence, Phrasing

10.21437/Interspeech.2012-631

1. Introduction

Each word has a canonical pronunciation – the way it would be pronounced if it were carefully articulated in isolation. However, words are frequently pronounced differently from their canonical pronunciation. The realized pronunciation can be significantly different due to coarticulation effects and other articulator co-ordination mistakes. A number of researchers have explored the role of prosodic variation on phonetic realization. Accented vowels are less affected by coarticulation while prosodic boundaries encourage more precise articulation. This impacts the realization of phrase final vowels and phrase initial consonants [1]. Phrase-final vowels and phrase-initial consonants are less reduced around phrase boundaries [2]. Borys et al. [3] found that when labeled data is available, prosodic information is *more* helpful for acoustic modeling than phonetic context is. Others have found that prominence and phrasing have a significant impact on articulation leading to altered acoustic realizations of phonetic units [4, 5, 6].

The simplest style of pronunciation model is a dictionary, where each lexical entry includes a single pronunciation represented as a string of phones. Using this model, the single canonical pronunciation is retrieved from the dictionary. This lookup-style pronunciation model can be augmented to allow for a single lexical entry to have multiple pronunciations. This can allow the model to accommodate a wider variety of realizations for a given word. However, not all pronunciations are equally likely. Dictionary models can be used as a probabilistic model by including a weight corresponding to $p(pr|w)$, the likelihood of pr, a pronunciation, given w the word identity. This allows increased flexibility, and can be easily trained from annotated corpora. More sophisticated pronunciation models calculate the likelihood of a pronunciation of a word based on its orthography, its canonical pronunciation, and other features including n-gram and part-of-speech tags.

In this work, we explore the incorporation of prosodic hypotheses in a weighted-dictionary pronunciation model. By considering a simpler model, we can focus our attention on the influence of prosody on pronunciation variability. We address the question of whether the variability in word pronunciations can be reduced through hypotheses of pitch accent and intonational phrase boundary locations. This type of model can be applied to improve automatic speech recognition (ASR) either through lattice and n-best rescoring or through direct incorporation into an ASR decoder.

In Section 2, we describe previous research on the relationship between prosody and pronunciation variation. We describe the Buckeye Corpus in Section 3. In Section 4, we present the pronunciation model and technique for predicting prosodic events. We discuss experimental results in Section 5. We conclude and describe future work in Section 6

2. Related Work

There is some related work on the incorporation of prosody into pronunciation models. Chen and Hasegawa-Johnson [7] describe how prosody can be incorporated into a Bayesian network to improve speech recognition. This work describes the incorporation of hypothesized prosodic labels for accent and intonational phrase boundaries in the acoustic, pronunciation and language models of a continuous speech recognizer. While the impact on pronunciation modeling is not explicitly addressed, the overall WER is reduced by 2.5% through the incorporation of prosodic variables.

Ostendorf et al. [8] investigated the use of prosody

conditioning in spontaneous speech recognition. In addition to incorporating prosodic information in acoustic modeling, this work directly incorporates acoustic/prosodic features into a statistical pronunciation model. This work is an example of "direct" modeling of prosody [9], where acoustic features, rather than symbolic representations of prosody are used by the model. The pronunciation model that this paper describes uses a trigram of baseform phone labels, part-of-speech labels, stress information and the location of the phone in the syllable. This work finds a small gain in pronunciation model perplexity due to the incorporation of prosody.

Bates and Ostendorf [10] investigated the use of F0, duration and energy in pronunciation modeling. This work found that incorporation of prosodic features in a pronunciation model that uses phonetic context and word-based features significantly improved the models. Finke and Waibel [11] investigated the incorporation of speaking mode information into a pronunciation model. This was accomplished in part by leveraging prosodic information to recognize different speaking rates and styles. Fosler-Lussier and Morgan also found that speaking rate modeling can lead to improvements in recognition accuracy [12].

3. Buckeye Corpus

The Buckeye Corpus [13] is a collection of autobiographical, sociolinguistic interviews. The material is drawn from forty speakers. The speech was collected in Columbus, Ohio, and to some degree represents regional pronunciation variation. Critically for this work, the Buckeye Corpus has been manually orthographically and phonemically transcribed. For each time aligned word, both the canonical (dictionary) pronunciation and realized pronunciation are available. The Buckeye Corpus includes 284,822 word tokens with a vocabulary of 9,568 unique words. By comparison this is much larger than other phonemically labeled corpora including TIMIT [14] with 6,300 words and the labeled subset of the Switchboard corpus (35,000) words [15].

4. Prosodic Pronunciation Model

To model the pronunciation variation in the Buckeye Corpus, we use a weighted-dictionary pronunciation model.

This model calculates $p(pr|w)$ where pr is a pronunciation phone sequence, and w is a word token based on a set of phonemically transcribed training data. We incorporate predictions of hypothesized prosodic events $\vec{e}$ by extending this model to $p(pr|w,\vec{e})$. We calculate these likelihoods based on observed counts in the training data. In the experiments reported in this paper we do not do any smoothing of the model to account for unobserved or infrequently observed words or pronunciations. This allows us to measure the rate at which unseen pronunciations and out-of-vocabulary words are produced. When calculating $p(pr|w,\vec{e})$, if there are zero observations of $w,\vec{e}$, we back off to the $p(pr|w)$ model. That is, when a word w has never been observed in the current prosodic context $\vec{e}$, we use the pronunciation model that does not include prosodic information.

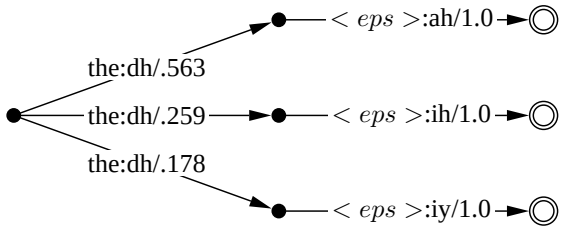

Figure 1: *An FST representation of the top three pronunciations of "the".*

This style of weighted-dictionary pronunciation model can be represented as an FST to facilitate composition with an ASR lattice. An FST representation of the model for the three most frequent Buckeye pronunciations of "the" appears in Figure 1.

In this work, we explore three prosodic events: the presence of a pitch accent (a), a following intonational phrase boundary (p_f), and a preceding intonational phrase boundary (p_b). We use AuToBI [16] to hypothesize the location of pitch accents and intonational phrase boundaries for every word in the Buckeye Corpus. The Buckeye Corpus does not include prosodic annotations. Therefore, we use detection models trained on the Boston Directions Corpus (BDC) [17] to generate prosodic hypotheses. The BDC includes speech from four non-professional speakers. Each speaker performed a series of nine direction giving tasks. This elicited spontaneous speech was subsequently transcribed manually, and speech errors were removed. Subjects later returned to the lab and read transcripts of their spontaneous monologues. While the content of the BDC material and the Buckeye Corpus different, both corpora include monologue speech produced by non-professional speakers.

AuToBI generates prosodic predictions using pitch, intensity, duration, pausing and spectral features. For both prediction tasks, features are normalized based on the speech file currently being processed. For pitch accent prediction an additional set of features based on context normalization are also used. Pitch accent detection is performed with Logistic Regression classifiers, while intonational phrase boundary detection uses AdaBoost with single split deception trees.

5. Experiments and Discussion

We use perplexity to evaluate the amount of pronunciation variability in the Buckeye Corpus. Perplexity is calculated as $2^{\frac{1}{N}\sum_n \log p(pr_n|w_n)}$ where N is the number of evaluated words, and w_n and pr_n are the n-th word and

its pronunciation. Lower perplexity indicates lower variability, and a better model.

We first evaluate the pronunciation model perplexity on the full Buckeye Corpus. This gives a baseline measure of the amount of pronunciation variation. We evaluate the impact of the introduction of each prosodic variable separately and combinations of prosodic variables. Some reduction of perplexity may be due to the fact that the prosodic variables are providing extra parameters to the pronunciation model. Therefore, we compare these perplexity scores to those obtained from a pronunciation model that includes a number of random binary features variables equal to the number of used prosodic variables. The perplexity results appear in Table 1. We find

Prosodic Features	Perplexity	Random Baseline
None	5.416	NA
a	5.099	5.305
p_f	5.119	5.305
p_p	5.197	5.305
a, p_f	4.823	5.265
a, p_f, p_p	4.619	5.246

Table 1: *Perplexity of prosodic pronunciation models on the full Buckeye Corpus*

that the inclusion of prosodic variables into the pronunciation model significantly reduces the model perplexity. While the introduction of random binary variable, does reduce pronunciation simply by adding extra parameters, the performance from the prosodic variables is better than the random variable baselines.

To see how well these results generalize across speakers, we perform leave-one-speaker-out cross validation of the perplexity results. In Table 2, we report the pronunciation model perplexity, as well as the frequency at which the model backed off to the non-prosodic pronunciation model, and the rate at which unseen words and unseen pronunciations appeared. We again find substan-

Prosodic Features	Perp.	Backoff Rate	Random Baseline
None	6.334	0.0	NA
a	6.133	.123	6.277
p_f	6.173	.123	6.277
p_p	6.223	.120	6.277
a, p_f	6.005	.141	6.234
a, p_f, p_p	5.921	.155	6.181

Table 2: *Perplexity of prosodic pronunciation models with leave-one-speaker-out cross-validation*

tial improvements to model perplexity through the introduction of prosodic variables. We note that this performance is based only on observed annotated pronunciations. This makes the model especially sensitive to out-of-vocabulary errors and zero-probability, unseen pronunciations. The perplexity measures only include those tokens for which the pronunciation has been seen at least once before. Uncounted in this performance measure are a 2.47% OOV rate. This indicates that 2.47% of word tokens in the Buckeye Corpus are spoken by a single speaker. A dictionary based pronunciation model will not be able to accommodate these OOV terms. Moreover, 8.38% of pronunciations are unobserved. This value counts the rate at which a particular pronunciation for a given lexical item only appears in the phonemic transcript of a single speaker.

The Buckeye Corpus represents a particularly diverse set of pronunciations. For example, 4533 of the 9568 words have more than one pronunciation. The word 'that' has more different phonetic realizations than any other word with 355 including /dh ae tq/ (1083 instances) and /n eh tq/ (25). A histogram of the number of pronunciations appears in the left subfigure in Figure 2. While most words have a unique pronunciation, this histogram has a long tail. The mean number of pronunciations per word is 3.78.

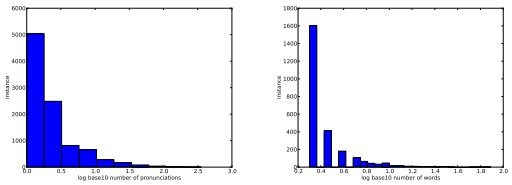

Figure 2: *A histograms of the number of distinction pronunciations annotated for each word and the number of words for each pronunciation*

Addressing the problem in the other direction, we find that of the 30,000 unique word pronunciations, 2,575 correspond to multiple words. Note that since word boundaries are available, this is a much simpler decoding task than the decoding of phone sequences to word sequences. The most ambiguous pronunciation is /ah/ which can correspond to 78 distinct words including "a", "uh", "you", "all", and "it". Some of this ambiguity is based on homophony; /s eh n t s/ is a pronunciation of "cents" "sense" and "since", /th eh r/ is a pronunciation of "there", "they're" and "their". Other cases are less predictable, including /d ao er/ as a pronunciation of "daughter", "dollar" and "door". A histogram of the number of words per pronunciation appears in the right subfigure in Figure 2; unambiguous pronunciations have been eliminated to more easily see the tail of the distribution.

We find that the pronunciation model perplexity decreases when treating adding an accent feature to the words. By inspecting the constructed models, we find that ACCENTED tokens have an average number of pronunciations of 3.32, while the DEACCENTED words have an average of 4.78 pronunciations. This is consistent with

the hypothesis that accenting leads to more careful and consistent pronunciation. On the other hand, deaccenting may lead to greater coarticulatory influence, less precise articulator timing and therefore a wider range of phonetic realizations. We also find that tokens *preceding* phrase boundaries have an average of 2.21 pronunciations while those that don't have an average of 3.67. Words that *follow* phrase boundaries have an average of 2.44 pronunciations while those that don't have an average of 3.66.

We also find that 9358 out of 9568 word types are realized with a hypothesized pitch accent, while only 2239 word tokens are predicted to be decaccented. This discrepancy suggests that words that words that are deaccented represent a more narrow class of words than words that are accented. Content words are accented between 60-75% of the time, while function words are more likely to be have a deaccented rate of 75-85% [18]. This discrepancy suggests a way distinguish function from content words based on their accent rate. In Figure 3, we plot the precision and recall of detecting function words by thresholding the deaccenting rate of each lexical item that occurs in the corpus more than three times. Though there is a correlation between accenting and function words, we find that this approach works less reliably than function word identification based on observation frequency. Here, we hypothesize that function words appear more frequently than content words. Additional work is necessary to understand how to use prosodic information for unsupervised word-classing.

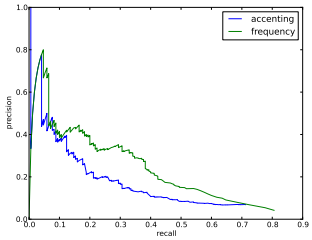

Figure 3: *Precision and Recall of function word detection using accent rate (blue) and frequency (green)*

6. Conclusions and Future Work

We find that prosodic hypotheses can be used to reduce pronunciation model perplexity. Words that precede or follow intonational phrase boundaries and accented words have less pronunciation variability than those that do not. This supports the hypotheses that the presence of these prosodic events leads to segmental productions that are more resistant to coarticulatory effects. We found some preliminary evidence that accent hypotheses can be used for word classing.

In the future work, we will incorporate these findings into ASR and Keyword search systems to determine the impact on word error rate. We will expand this evaluation to include the phonetically annotated portion of Switchboard. Application of these findings into a more sophisticated pronunciation model will serve to reduce the overall model perplexity and also determine the relative importance of prosody to n-gram and pronunciation distance measures. We expect to investigate the use of prosodic hypotheses in pronunciation models of languages other than English.

7. References

[1] T. Cho, *The Effects of Prosody on Articulation in English*. Routledge, 2002.

[2] P. Fougeron and P. Keating, "Articulatory strengthening at edges of prosodic domains," *Journal of the Acoustical Society of America*, vol. 101, no. 6, pp. 3728–3740, 1997.

[3] S. Borys, M. Hasegawa-Johnson, J. Cole, and A. Cohen, "Modeling and recognition of phonetic and prosodic factors for improvements to acoustic speech recognition models," in *Interspeech*, 2004.

[4] C. W. Wightman, S. Shattuck-hufnagel, M. Ostendorf, and P. Price, "Segmental durations in the vicinity of prosodic phrase boundaries," *JASA*, vol. 91, 1992.

[5] K. D. Jong, "The supraglotal articulation of prominence in english: Linguistic stress as localized hyperarticulation," *Journal of the Acoustic Society of America*, vol. 97, no. 1, pp. 491–504, 1995.

[6] J. Edwards, M. Beckman, and J. Fletcher, "The articulatory kinematics of final lengthening," *Journal of the Acoustical Society of America*, vol. 89, no. 1, pp. 369–382, 1991.

[7] K. Chen and M. Hasegawa-Johnson, "How prosody improves word recognition," in *Speech Prosody*, 2004.

[8] M. Ostendorf, I. Shafran, and R. Bates, "Prosody models for conversational speech recognition," in *Proc. of the 2nd Plenary Meeting and Symposium on Prosody and Speech Processing*, 2003, pp. 147–154.

[9] E. Shriberg and A. Stolcke, "Direct modeling of prosody: an overview of applications in automatic speech processing," in *Speech Prosody*, 2004, pp. 575–582.

[10] R. Bates and M. Ostendorf, "Modeling pronunciation variation in conversational speech using prosody," in *ISCA Tutorial and Research Workshop on Pronunciation Modeling and Lexicon Adaptation for Spoken Language*, 2002.

[11] M. Finke and A. Waibel, "Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition," in *Eurospeech*, 1997, pp. 2379–2382.

[12] E. Fosler-Lussier and N. Morgan, "Effects of speaking rate and word frequency on conversational pronunciations," in *Proc. of ESCA Pronunciation Modelling Workshop*, 1998, pp. 35–40.

[13] M. A. Pitt, L. Dilley, K. Johnson, S. Kiesling, W. Raymond, E. Hume, and E. Fosler-Lussier, "Buckeye corpus of conversational speech (2nd release)," [www.buckeyecorpus.osu.edu], 2007.

[14] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgrena, and V. Zue, "Timit acoustic-phonetic continuous speech corpus," Linguistic Data Consortium, 1993.

[15] S. Greenberg, "The switchboard transcription project," in *Proceedings of the Large Vocabulary Continuous Speech Recognition Summer Research Workshop*, Baltimore, Maryland, USA, April 1996.

[16] A. Rosenberg, "Autobi – a tool for automatic tobi annotation," in *Interspeech*, 2010.

[17] C. Nakatani, J. Hirschberg, and B. Grosz, "Discourse structure in spoken language: Studies on speech corpora," in *AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, 1995.

[18] A. Rosenberg, "Automatic detection and classification of prosodic events," Ph.D. dissertation, Columbia University, 2009.