How Prosody Improves Word Recognition

Ken Chen and Mark Hasegawa-Johnson

Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign, Urbana, IL 61801
{kenchen; jhasegaw}@uiuc.edu

Abstract

Prosody has been traditionally regarded as useless for word recognition. In this paper, we provide a schematic view describing how prosody can help word recognition. We provide our view in terms of a Bayesian network that models the stochastic dependence among acoustic observation, word, prosody, syntax and meaning, and an information-theoretic analysis proving that the mutual information between acoustic observation and correct word hypotheses improves if prosody is jointly modeled with word in a prosody dependent speech recognition framework. We also report our experiment on Radio News Corpus in which prosody has improved word recognition accuracy by 2.5%.

1. Introduction

Prosody has been traditionally regarded as useless for word recognition since acoustic-prosodic features are mostly suprasegmental and are only weakly dependent on phonetic models. The only prosodic feature that has been widely used in speech recognizer is the normalized energy. Various attempts have been made to incorporate duration into phonetic or word models, but only small improvement has been achieved when duration dependent models are applied to large scale continuous speech recognition. There are also studies that attempted to incorporate pitch into speech recognizer either by conditioning cepstral observations on pitch for normalization purpose, or by including pitch as auxiliary variable to create pitch dependent acoustic model [1]. The improvement reported by these attempts is small and there are no explicit prosody knowledge built into these systems.

On the other hand, due to the dependence of prosody on high-level linguistic units such as disfluency, syntax, dialog act, topic, meaning and emotion, prosody has been successfully used to disambiguate syntactically distinct sentences with identical phoneme strings, infer punctuation of a recognized text, segment speech into sentences and topics, recognize the dialog act labels [2], and detect speech disfluencies.

Can prosody ever help word recognition? Linguistic study has confirmed that humans are able to understand the content with lower cognitive load and higher accuracy while listening to natural prosody, as opposed to monotone or foreign prosody [3]. This suggests that it is possible to utilize prosody to improve automatic word recognition. In section 2, we describe the complex relationship among prosody and other linguistic units using a Bayesian network and suggest possible ways of using prosody to improve word recognition. Section 3 proposes our prosody dependent speech recognition framework and provides an information-theoretic analysis proving that word recognition can be improved when prosody dependence is incorporated in both acoustic model and language model. Section 4 briefly describes our experiments and results on Radio News Corpus, as an experimental proof to our theoretical analysis. Conclusions are given in section 5.

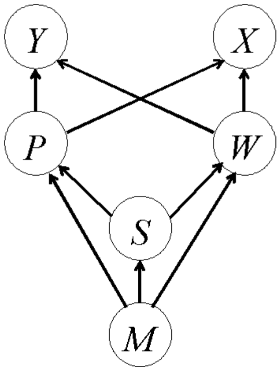

Figure 1: *A Bayesian network representing the complex relationship among the acoustic-phonetic features (X), acoustic-prosodic features (Y), word sequence (W), prosody sequence (P), syntax sequence (S) and meaning (M) of an utterance.*

2. Bayesian network

The complex relationship among prosody and other linguistic variables for a given utterance can be analyzed in terms of a Bayesian network as depicted in Fig. 1, where we use P to denote a sequence of prosody labels, one associated with each word in the word sequence W, describing the prosodic status (e.g., the accentuation and the lengthening) of each word, S represents a sequence of labels describing the syntactical role of each word (e.g., parts-of-speech tags can be used as a subset of S), M represents the meaning of the utterance that may affect the distribution of S, W and P. M is abstract in the sense that it represents not only the literal meaning of the utterance but also the high-level contextual meaning dependent on the neighboring utterances and other high level linguistic variables such as dialog act, topic and emotion. X is the acoustic-phonetic observation sequence sampled at either frame or segmental level and Y is the acoustic-prosodic observation sequence sampled at syllable or word level.

As shown in Fig. 1, X is dependent on both W and P. Prosody does affect the acoustic realization of words. For example, unaccented vowels tend to be centralized or even deleted in a function word, accented vowels tend to be longer and less subject to coarticulatory variation [4]; accented consonants are produced with greater closure duration, greater linguopalatal contact, longer voice onset time, and greater burst amplitude [5]. These phenomena can be modeled in ASR by introducing a prosody dependent pronunciation model $p(Q, H|W, P)$ and creating a prosody dependent acoustic model $p(o_l|q_l, h_l)$ (i.e., conditioning the PDFs in conventional monophone or triphone models on prosody variable h_l), where $Q = (q_1, \ldots, q_L)$ is a sequence of sub-word units, typically allophones dependent on phonetic context, $H = (h_1, \ldots, h_L)$ is a sequence of discrete "hidden mode" vectors describing the prosodic states of each allophone, and o_l is the acoustic observations (including both phonetic and prosodic observations) over the allophone q_l.

Y is also dependent on both W and P. Y depends on W in the sense that the location and the properties of prosodic events are constrained by the pronunciation of words. For example, in most cases, only primary lexical stressed syllables of an accented word are accented and only the rhyme of the last syllable in the words preceding prosodic boundaries are relatively lengthened [6].

P and W are mutually dependent. The dependence of P over W has been illustrated by Kompe [7] who found that prosody can be accurately predicted from word strings given enough training data. On the other hand, W is restricted when P is given because P can only be produced by certain word sequences. The mutual dependence of W and P can be more clearly understood through their dependence on S. The dependence of W on S is well-established. The dependence of P on S has been proved empirically by Arnfield in a corpus-based study [8] in which he claimed that although differing prosodies are possible for a fixed syntax, the syntax of an utterance can be used to generate an underlying "baseline" prosody regardless of actual words, semantics or context. The dependence of W on S has been successfully utilized to create factorial language models in which parts-of-speech are used as word categories [9]. Similarly, the dependence of P on S has been utilized to create factorial prosodic language models [10]. P and W can be assumed to be conditionally independent given S and M: $p(P|W, S, M) \approx p(P|S, M)$ and $p(W|P, S, M) \approx p(W|S, M)$.

W, P and S are all dependent on M. The dependence of P over M can be either strong or weak depending mainly on the speech mode. In normal conversational speech, P is weakly dependent on M except for the cases when there are syntactical ambiguities or there are high level information (such as emphasis, contrast, attitude and emotion) to convey.

Prosody can be used to help word recognition in at least three ways. First, it can be used to improve the accuracy of the acoustic model due to the dependence of X over both W and P. Second, since P and W are strongly dependent on each other, they can be modeled as a single variable and the acoustic-phonetic observation and the acoustic-prosodic observation can be combined into one observation stream: $O = [X, Y]$. This is so called prosody dependent speech recognition which we will discuss in details in section 3. In our current system, S is not explicitly modeled (due to the small size of the corpus). Instead, it is used to reduce the conditional entropy of $p(W, P)$ through factorial based approaches. Heeman has proposed to explicitly model S in the language model but he has assumed that O is independent of P and S in his acoustic model [9] (and he has not integrated into O the acoustic-prosodic observation Y). The third way is to use the knowledge of P to infer the status of M which can then be used to reduce the conditional entropy of W. This approach has been illustrated by Taylor [2] who conditioned the language model on dialog act labels that can be inferred from prosody. This way of using prosody is highly task-dependent and is limited by the availability of explicit representation of M.

3. Prosody dependent speech recognition

The task of speech recognition, given a sequence of observed acoustic feature vectors $O = (o_1, \ldots, o_T)$, is to find the sequence of word labels $W = (w_1, \ldots, w_M)$ that maximizes the recognition probability:

$$[\tilde{W}] = \arg\max p(O, W), \tag{1}$$

The mutual dependence between word and prosody has motivated us to model them as a joint unit:

$$[\tilde{W}] = \arg\max p(O|Q, H)p(Q, H|W, P)p(W, P) \tag{2}$$

where $p(O|Q, H)$ is a prosody-dependent acoustic model, $p(Q, H|W, P)$ is a prosody-dependent pronunciation model, and $p(W, P)$ is a prosody-dependent language model. The combination $[w_m, p_m]$ is called a prosody-dependent word label, the combination $[q_l, h_l]$ is called a prosody-dependent allophone label,

Let a prosody-dependent allophone model be defined as an HMM whose states are conditioned on both phoneme label q_l and prosodic state h_l. Assume that a prosody-dependent pronunciation model may be pre-compiled so that each prosody-dependent word label $[w_m, p_m]$ corresponds to a unique hidden Markov model, created by concatenating an appropriate sequence of prosody-dependent allophone models. Let the prosody dependent language model be defined to be any standard language model (this paper will use bigram models) describing the probability of $[w_m, p_m]$ given the history $[w_1, p_1, \ldots, w_{m-1}, p_{m-1}]$. The average modeled mutual information between the true word hypothesis W_T and the acoustic observation O may be defined as:

$$I(O; W_T) = E_{W_T,O}\left\{\log \frac{p(O, W_T)}{p(O)p(W_T)}\right\}, \tag{3}$$

where the expectation is computed over the true joint distribution of W_T and O, but the probabilities in the fraction are modeled probabilities; thus $I(O; W_T)$ is a measure of the quality of the PDF model $p(O, W_T)$. Suppose that $p(W_T)$ in (3) is defined to be the true probability of W_T, so that only the terms $p(O)$ and $p(O, W_T)$ depend on the quality of the speech recognition model. Under this definition, the quantity $I(O; W_T)$ is related by a constant to the model discriminant function $\Phi(O; W_T)$ [11], defined as:

$$\begin{aligned}
&\Phi(O; W_T)\\
&\quad= \quad E_{W_T,O}\left\{\log p(W_T|O)\right\}\\
&\quad= \quad E_{W_T,O}\left\{\log \frac{p(O, W_T)}{\sum_{\hat{W}} p(O, \hat{W})}\right\}\\
&\quad= \quad -E_{W_T,O}\left\{\log\left(\sum_i \eta_i\right)\right\},
\end{aligned} \tag{4}$$

where

$$\eta_i = \frac{p(O, \hat{W}_i)}{p(O, W_T)} = \frac{p(O|\hat{W}_i)}{p(O|W_T)} \times \frac{p(\hat{W}_i)}{p(W_T)}, \tag{5}$$

which is the likelihood ratio comparing the ith word sequence hypothesis $\hat{W}_i$ to the true word sequence W_T.

The discriminant function of a prosody dependent recognizer can be represented as

$$\Phi_P(O; W_T) = -E_{W_T,O}\left\{\log\left(\sum_i \acute{\eta}_i\right)\right\} \tag{6}$$

where

$$\begin{aligned} \acute{\eta}_i &= \frac{\max_{\hat{P}} p(O, \hat{W}_i, \hat{P})}{\max_{\hat{P}} p(O, W_T, \hat{P})}, \\ &= \frac{p(O|\hat{W}_i, \hat{P}_i)}{p(O|W_T, P_T)} \times \frac{p(\hat{W}_i, \hat{P}_i)}{p(W_T, P_T)}, \end{aligned} \tag{7}$$

P_T is the prosody sequence that maximizes $p(O, W_T, \hat{P})$, and $\hat{P}_i$ is the prosody hypothesis that maximizes $p(O, \hat{W}_i, \hat{P})$.

The objective of prosody-dependent speech recognition in this paper is to create prosody-dependent speech recognition models such that $\Phi_P(O; W_T) > \Phi(O; W_T)$, thus increasing the modeled probability of the correct word sequence given the observation. From (4) and (6), $\Phi_P(O; W_T) > \Phi(O; W_T)$ if

$$E_{W_T,O}\left\{\log\left(\frac{\sum_i \acute{\eta}_i}{\sum_i \eta_i}\right)\right\} < 0 \tag{8}$$

Equation (8) expresses the condition under which prosody-dependent speech recognition increases the modeled mutual information $I(O; W_T)$. In order to guide the design and interpretation of experiments in the field of prosody-dependent speech recognition, it is valuable to spend some time trying to express the meaning of equation (8) in words. Loosely speaking, equation (8) claims that modeled mutual information improves if $\acute{\eta}_i < \eta_i$ for most combinations of W_T and $\hat{W}_i$, where the word "most" is quantified by the expectation over W_T of the log ratio of sums over $\hat{W}_i$. Re-arranging terms, the condition $\acute{\eta}_i < \eta_i$ may be written:

$$\left(\frac{p(P_T|W_T)}{p(\hat{P}_i|\hat{W}_i)}\right)\left(\frac{p(O, W_T|P_T)/p(O, W_T)}{p(O, \hat{W}_i|\hat{P}_i)/p(O, \hat{W}_i)}\right) > 1 \tag{9}$$

Equation (9) expresses the fraction $\eta_i/\acute{\eta}_i$ as the product of two terms. The first term on the left expresses the improvement, due to prosody, in the selectivity of the language model. This term is positive, for example, when the true word sequence is uttered with a highly predictable prosodic pattern, thus $p(P_T|W_T) > p(\hat{P}_i|\hat{W}_i)$. The second term on the left expresses the improvement, due to prosody, in the selectivity of the acoustic model. This term is positive, for example, when the observation sequence O is better explained by the true prosody P_T than by any false prosody $\hat{P}_i$. The meaning of Equation (8) may therefore be explained in the following words: $\Phi_P(O; W_T) > \Phi(O; W_T)$ if, most of the time, the correct prosodic sequence is well predicted by the word transcription, and the acoustic observation is well predicted by the prosody. Note that it is possible for a prosody-dependent speech recognizer to result in improved word recognition accuracy even if the acoustic model and the language model do not separately lead to improvements. Even if prosody does not improve the recognition of words in isolation, the likelihood of the correct sentence-level transcription may be improved by a language model that correctly predicts prosody from the word string, and an acoustic model that correctly predicts the acoustic observations from the prosody.

The selectivity of the language model may be maximized by modeling only those prosodic labels that are most predictable from word sequence statistics. In this paper, prosodic labeling will include intonational phrase boundaries and phrasal pitch accent. Previous research [7] has shown that both intonational phrase boundaries and phrasal pitch accent are well predicted by N-gram word sequence statistics.

The selectivity of the acoustic model may be maximized by selectively modeling only those acoustic features whose distributions are well predicted by prosodic labeling. Papers in acoustic phonetics suggest that talker-normalized fundamental frequency (f_0) is well predicted by the location of pitch accents [12], while normalized phoneme duration is well predicted by the location of intonational phrase boundaries [6]. Prosody-dependent modification of the acoustic-phonetic features (e.g., MFCC) has been described as a reliable effect in the case of some phonemes but not all phonemes [5], thus prosody-dependent modification of the distribution of MFCCs will be modeled only for an empirically selected subset of phonemes.

4. Experiments and results

We have trained and tested our prosody dependent speech recognizer on the Boston University Radio News Corpus [13]. In this corpus, a majority of paragraphs are annotated with the orthographic transcription, phone alignments, part-of-speech tags and prosodic labels. The part-of-speech tags used in this corpus are the same as those used in the Penn Treebank. Part-of-speech labeling is carried out automatically using the BBN tagger.

The prosodic labeling system represents prosodic phrasing, phrasal prominence and boundary tones, using the Tones and Break Indices (ToBI) system for American English. The ToBI system labels pitch accent tones, phrase boundary tones, and prosodic phrase break indices. Break indices indicate the degree of decoupling between each pair of words; intonational phrase boundaries are marked by a break index of 4 or higher. Tone labels indicate phrase boundary tones and pitch accents, constructed from the three basic elements H, L, and !H which represent high tone, low tone, and high tone followed by pitch downstep, respectively. In the experiments we reported in this paper, the original ToBI labels are simplified: pitch accents are only distinguished by presence versus absence, word boundaries are only distinguished by intonational phrase boundary versus non-intonational-phrase-boundary. Applying this simplification, we create prosody dependent word transcriptions in which a word can only have four possible prosodic variations: unaccented phrase medial ("um"), accented phrase medial ("am"), unaccented phrase final ("uf") and accented phrase final ("af").

The prosodically labeled data used in our experiments consist of 300 utterances, 24944 words (about 3 hours of speech sampled at 16Khz) read by five professional announcers (3 female, 2 male) containing a vocabulary of 3777 words. Training and test sets are formed by randomly selecting 85% of the utterances for training, 5% of the utterances for development test and the remaining 10% for testing (2503 words). Two acoustic models are used in this experiment: a prosody independent acoustic model API and a prosody dependent acoustic model APD. All phonemes in API and APD are modeled by HMMs consisting of 3 states with no skips. Within each state, a 3 mixture Gaussian model is used to model the probability density of a 32-dimensional acoustic-phonetic feature stream consisting of 15 MFCCs, energy and their deltas. The allophone models in APD contain an additional one-dimensional Gaussian acoustic-prosodic observation PDF which is used to model the probability density of a nonlinearly-transformed pitch stream [14]. API contains monophone models adopted from the standard SPHINX set [15] and is unable to detect any prosody related acoustic effects. APD contains a set of prosody dependent allophones constructed from API by splitting the monophones into allophones according to a four-way prosodic distinction (unaccented medial, accented medial, unaccented final, accented

Table 1: *Percent word, accent and intonational phrase boundary recognition accuracy for recognizers RII, RID, RDM, and RDC.*

	RII	**RID**	**RDM**	**RDC**
AM	API	APD	APD	APD
LM	LPI	LPI	LPDM	LPDC
Word	75.85	76.02	77.29	78.27
Accent	56.07	56.07	79.59	80.26
IPB	84.97	84.97	85.06	86.62

final): each monophone in API has 4 prosody dependent allophonic variants in APD. Allophone models in APD that are split from the same monophone share a single tied acoustic-phonetic observation PDF, but each allophone distinctly models the state transition probabilities and the acoustic-prosodic observation PDF. The APD allophones are therefore able to detect two of the most salient prosody induced acoustic effects: the preboundary lengthening, and the pitch excursion over the accented phonemes. The parameter count of the acoustic-phonetic observation PDF (195 parameters per state) is much larger than the parameter count of the acoustic-prosodic observation PDF (2 parameters per state) or the transition probabilities (1 parameter per state); since the acoustic-phonetic parameters are shared by all allophones of a given monophone, the total parameter count of the APD model set is only about 6% larger than the parameter count of API.

Three language models are trained from the same training set: a standard prosody independent backoff bigram language model LPI, a prosody dependent backoff bigram language model LPDM computed using the prosody dependent ngram count, a prosody dependent factorial backoff bigram language model LPDC as reported in [10] in which parts-of-speech are used as word classes to reduce the language model perplexity.

Four recognizers are tested: a standard prosody independent recognizer RII using API and LPI, a semi-prosody independent recognizer RID using APD and LPI, a prosody dependent recognizer RDM using APD and LPDM, and a prosody dependent recognizer RDC using APD plus LPDC. The word recognition accuracy, accent recognition accuracy and intonational phrase boundary recognition accuracy of these recognizers over the same training and test set are reported in Table 1.

Overall, the prosody dependent speech recognizers have significantly improved the word recognition accuracy (WRA) over the prosody independent speech recognizers. RDM has improved the word recognition accuracy by 1.4% over RII and 1.2% over RID. RDC has further improved the WRA by 1% over RDM, apparently benefitting from the improved prosody dependent language model LPDC. The pitch accent recognition accuracy (ARA) and the intonational phrase boundary recognition accuracy (BRA) are also significantly improved. Since RII and RID classify every word as unaccented and every word boundary as phrase-medial, the ARA and BRA listed in RII and RID are the chance levels. RDC has achieved a significant improvement of ARA and BRA: 24.2% and 1.7% repectively above the chance levels.

5. Conclusions

In this paper, we have schematically analyzed the stochastic dependence among acoustic observations, word, prosody, syntax and meaning using a Bayesian network and suggested possible ways of using prosody to improve word recognition. We have proposed a prosody dependent speech recognition framework and provided an information-theoretic analysis proving that the mutual information between acoustic observation and correct word hypotheses improves if prosody is jointly modeled with word. The word recognition results and the prosody recognition results on Radio News Corpus have shown that prosody can improved word recognition accuracy by as large as 2.5%.

6. References

[1] Stephenson, T.A.; Mathew, M.; Bourlard, H., 2001. Modeling auxiliary information in Bayesian network based ASR. *Eurospeech 2001.*

[2] Taylor, P.; King, S.; Isard, S.; Wright, H.; Kowtko, J., 1997. Using intonation to constrain language models in speech recognition. *Eurospeech 1997.*

[3] Hahn, L., 1999. Native speakers' reactions to non-native stress in English discourse. Ph.D. dissertation, University of Illinois at Urbana-Champaign.

[4] Cho, T., 2001. Effects of prosody on articulation in English. Ph.D. dissertation, University of California at Los Angeles.

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

[6] 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, 1707-1717.

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

[8] Arnfield, S., 1994. Prosody and syntax in corpus based analysis of spoken English. Ph.D. dissertation, University of Leeds.

[9] Heeman, P.; Allen, J., 1999. Speech repairs, intonational phrases and discourse markers: modeling speakers' utterances in spoken dialog. *Computational Linguistics*, vol. 25, no. 4.

[10] Chen, K.; Hasegawa-Johnson, M., 2003. Improving the robustness of prosody dependent language modeling based on prosody syntax dependence. *IEEE ASRU 2003*.

[11] Normandin, Y.; Morgera, S. D., 1991. An improved MMIE training algorithm for speaker-independent, small vocabulary, continuous speech recognition. *IEEE ICASSP'91*, vol.1, 537-540.

[12] Beckman, M. E.; Pierrehumbert, J., 1986. Intonational structure in Japanese and English. *Phonology Yearbook*, vol. 3, 255-309.

[13] Ostendorf, M.; Price, P. J.; Shattuck-Hufnagel, S., 1995. *The Boston University Radio News Corpus.* Linguistic Data Consortium.

[14] Kim, S.; Hasegawa-Johnson, M.; Chen, K., 2004. Automatic recognition of pitch movements using time-delay recursive neural network. *IEEE Signal Processing Letter*, in press.

[15] Lee, K. F., 1990. Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition. *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 38, No. 4, 599-609.