A Maximum Likelihood Prosody Recognizer

Ken Chen, Mark Hasegawa-Johnson, Aaron Cohen and Jennifer Cole

Department of Electrical and Computer Engineering and Department of Linguistics
University of Illinois at Urbana-Champaign, U.S.A.
{kenchen; jhasegaw; ascohen; jscole}@uiuc.edu

Abstract

Automatic prosody recognition (APR) is very importance for automatic speech understanding. In this paper, we propose a maximum likelihood prosody recognizer consisting of a GMM-based acoustic model that models the distribution of the phone-level acoustic-prosodic observations (pitch, duration and energy) and a ANN-based language model that models the word-level stochastic dependence between prosody and syntax. Our experiments on the Radio News Corpus show that our recognizer is able to achieve 84% pitch accent recognition accuracy and 93% intonational phrase boundary (IPB) recognition accuracy in a leave-one-speaker-out task which has exceeded previous reported results on the same corpus. The same recognizer is tested on a subset of switchboard corpus. The accuracies are degraded but still significantly better than the chance levels.

1. Introduction

Prosody refers to the suprasegmental features of natural speech (such as rhythm and intonation) that are used to convey linguistic and paralinguistic information (such as emphasis, intention, attitude and emotion). Prosody affects the acoustic realization of speech at phonetic, syllabic, lexical and word level. At phonetic level, prosody affects the acoustic realization of phonemes. For example, accented vowels tend to be longer and less subject to coarticulatory variation [1], while accented consonants are produced with greater closure duration, greater linguopalatal contact and longer voice onset time. At syllabic level, prosody manifests through distinctive pitch and intensity movements and durational variation (e.g., pitch accents, boundary tones) under the constraints of the lexical stress patterns, which are relatively fixed as intrinsic properties of words. However, in the cases of emphatic accents (where all syllables of a word are accented) and contrastive accents (where accentuations are realized on non-primary-lexical stressed syllables), the stress assignment is determined directly by prosody other than lexicon. At word level, prosody manifests through phrasal prominences and meaningful uses of breaks and pauses under the constraints of syntax and other high-level linguistic variables such as pragmatics, semantics and emotion.

The correct recognition of prosody not only requires correct prosody recognition models at these different levels, but also requires a model that imposes the high-level linguistic constraints. Wightman et al. [2] proposed an automatic prosody recognition system that detects prosody at syllable-level. In their system, decision-tree models are trained to calculate the posterior probability of syllable-level prosody labels given the syllable-timed acoustic features. This recognizer lacks a model that imposes the high-level linguistic constraints and assumes that prosody can be determined completely from their syllabic acoustic observations and pre-compiled lexical stress information. Nevertheless, it is successful on labeling pitch accents on the Radio News Corpus [3] with 84% accuracy on accent presence/absence prediction, about 30% higher than the estimated chance level. However, it does not perform well on intonational phrase boundary (IPB) detection: IPB recognition accuracy is only 71%, 12% below the estimated chance level. The low IPB recognition accuracy is mainly due to the insufficient acoustic statistics at intonational phrase boundaries. Unlike the acoustic-phonetic features, the syllabic acoustic-prosodic features are intrinsically highly variable not only in their strength (amplitude, shape, duration) but also in their time alignment with the syllables (e.g., the peak or valley of the pitch contour may occur in the syllables preceding or succeding the accented syllable). In addition, they often suffer from both inter-speaker difference (e.g., female speakers usually use more expressive prosody than male speakers) and intra-speaker difference (e.g., a speaker can use different prosody for the same word strings in different context). In fact, determining prosody at syllable level from given acoustic context and lexical constraints is not only difficult for machines but also difficult for human labelers. While listening to the speech segments, human labelers often utilizes high-level linguistic constraints to predict prosody. For example, human labelers can use the fact that function words such as “of”, “the” are rarely accented to increase their labeling accuracy and speed.

The dependence of prosody over high-level linguistic variables has been applied in speech synthesis to assign prosody from text. Although it is generally believed that syntactic, semantic and discourse/pragmatic factors are all involved in prosody decision, such labeling relies primarily on syntactic analysis due to the difficulty in representing and extracting high-level linguistic information (the discourse, pragmatic and semantical information) from text. Hirschberg [4] has proposed a decision-tree based system that has achieved 82.4% speaker dependent accent labeling accuracy on Radio News, a large improvement over early systems that labels prosody based on function word versus content word distinction. Hirschberg’s result is important because it shows that it is possible to accurately predict prosody from syntax. In another corpus-based study, Arnfield [5] has claimed, after his bigram models predicted prosodic stress from parts-of-speech (POS) with 91% accuracy, 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. Similar results have been achieved by Ross [6], whose system predicts ToBI [7] style prosody labels from text with 82.5% word-level accent presence/absence accuracy. Ross’s decision-tree based system is different from Hirschberg’s in that it assigns prosody at syllable level instead of at word level and requires pre-generated prosodic phrase structure as input.

Even though the importance of syntax in predicting prosody has been recognized in designing these previous systems, the syntactic information contained in the text are not fully utilized: these systems either used small POS set (only 8 POS categories in [4] [6]) due to the limitation in their decision-tree algorithm, or included only small POS context (unigram in [4] [6] and bigram in [5]).

Kompe [8] has proposed another prosody recognition system that uses neural network for the acoustic-prosodic modeling of segmental prosody and a polygram model for the syntactic-prosodic modeling of word-wise prosody. The polygram model that he uses computes the probability of a prosody label p_l given the surrounding n words: $p(p_l|w_{l-n+2}, w_{l-n+3}, \ldots, w_{l+n-1})$. Kompe's system has achieved 95% IPB recognition rate for his prosodic-syntactical M labels, labels that are deterministically transformed from syntactic clause boundaries (based on a set of empirical rules) but better correlate with prosodic phrase boundaries than syntactic boundaries. Kompe's syntactical-prosodic model would be ideal given a large amount of training data. In practice, conditioning prosody on word strings creates problems of data-sparseness especially for small-sized corpora. Despite this disadvantage, Kompe's result suggests the potential advantage of modeling the dependence of prosody over large context ($n > 3$) and relatively large variety of word categories other than the over-simplified POS classes. Rather than conditioning prosody on word strings, conditioning prosody on their syntactical representation (e.g., parts-of-speech) can effectively reduce the entropy of the syntactic-prosodic models [9].

Motivate by these results, we propose to build a prosody recognizer that effectively detects acoustic-prosody cues and imposes syntactical constraints. In section 2, we formulate our system in a maximum-likelihood estimation framework, similar in appearance to canonical automatic speech recognizers. Section 3 describes the acoustic features and syntactical features that are used in our experiments. Section 4 reports the experiments and results, and conclusions are given in section 5.

2. Method

Let $W = (w_1, \ldots, w_L)$ be the word sequence, $P = (p_1, \ldots, p_L)$ the prosody sequence of an utterance. The task of prosody recognition is to find the optimal prosody sequence $\tilde{P}$ that maximizes the recognition probability:

$$\begin{aligned} [\tilde{P}] = & \arg\max_P p(Y, W), \\ = & \arg\max_P p(Y|W, P)p(P|W), \\ = & \arg\max_P \prod_{l=1}^{L} p(y_l|w_l, p_l)p(p_l|\phi_l(W))^{\gamma}, \end{aligned} \quad (1)$$

where $Y = (Y_1, \ldots, Y_L)$ is a sequence of L word-wise acoustic-prosodic features and $\phi_l(W)$ is a function that extracts all the information in W that affects the prediction of p_l. Assuming the dependence of prosody on word strings is localized in a window of n words and is described by the syntactical roles of the words (primarily parts-of-speech) instead of the words themselves, then:

$$\phi_l(W) = (s_{l-n+2}, s_{l-n+3}, \ldots, s_{l+n-1}) \quad (2)$$

where s_l represents the syntactic information contained in w_l that affects the prediction of p_l. In general, s_l can include all possible information one could obtain from the text analysis (including semantic information). Parts-of-speech is shown to be most useful, but other type of information such as the location of syntactic boundaries is also helpful. The language model probability has been raised by a power of γ, a constant that can be used to adjust the weighting between the language model and the acoustic model.

The probability $p(y_l|w_l, p_l)$ in equation (1) can be further expanded to syllable or phoneme level:

$$\begin{aligned} & p(y_l|w_l, p_l) \\ & = \sum_{Q_l, H_l} \prod_{(q_{n_l}, h_{n_l}) \in (Q_l, H_l)} p(y_{n_l}|q_{n_l}, h_{n_l}) p(Q_l, H_l|w_l, p_l) \end{aligned} \quad (3)$$

where $p(Q_l, H_l|w_l, p_l)$ is a pronunciation model that computes the probability of a phoneme string $Q = (q_1, \ldots, q_{n_l}, \ldots, q_{N_l})$ and the accompanying phoneme-level prosody string $H = (h_1, \ldots, h_{n_l}, \ldots, h_{N_l})$ given prosody dependent word token (w_l, p_l), and y_{n_l} represents the acoustic-prosodic observations over the allophone (q_{n_l}, h_{n_l}). Assuming that prosody does not change the assignment of lexical stress, all the pronunciation information can be pre-compiled and loaded in before recognition starts. Note that the lexical stress information can be conveniently expressed in the pronunciation model, so does the prosody induced pronunciation variation (different pronunciation of a word under difference prosody). An example is given below for the word "above":

- above: ax b ah v
- above!: ax b! ah! v!
- aboveB4: ax b ahB4 vB4
- above!B4: ax b! ah!B4 v!B4

In the above example, we used postfix "!" to label the pitch accent, and "B4" to label the words and phonemes that are under the influence of intonational phrase boundaries. Since in most cases, only the primary stressed syllable in an accented word is accented, the "!" label can only be attached to the phonemes in the primary lexical stressed syllable. "B4" can only be attached to phonemes in the last rhyme because it has been shown that preboundary lengthening only happen in the rhyme of the last syllable. Since a prosody dependent word token (w_l, p_l) may have multiple pronunciations, a summation over Q_l, H_l is included in equation (3) to sum up all possible lexical entries for (w_l, p_l).

The language model $p(p_l|\phi_l(W))$ has been modeled by a multilayer perceptron (MLP) where the output of the MLP is used to compute the posterior probability of p_l given the syntactical information $\phi_l(W)$:

$$p(p_l = i|\phi_l(W)) = \frac{g_i(\phi_l(W))}{\sum_i g_i(\phi_l(W))}. \quad (4)$$

where $g_i(\cdot)$ is the i^{th} output of the MLP. Since we have chosen $\phi_l(W)$ such that it contains syntactical information from a fixed window of n words surrounding p_l, the number of input nodes is always fixed for each l. The number of output nodes is determined by the variety of prosody that is modeled at word level. In this paper, we chose to model only 4 possible prosody patterns for each word: unaccented phrase-medial, unaccented phrase-final, accented phrase-medial and accented phrase-final. This set of prosody labels is the same as what has been used in [2].

The acoustic model $p(y|q, h)$ is trained using standard EM algorithm and The MLP-based syntactic prosodic model is trained using standard error back-propagation algorithm.

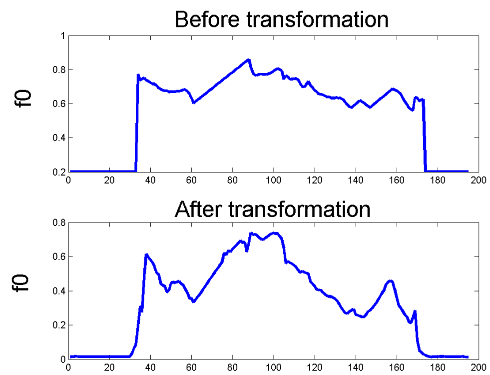

Figure 1: The non-linear pitch transformation of the utterance "ANd, FIve of them will be JEWish", where capital letters denote the accented syllable onsets.

3. Features

3.1. Acoustic features

The primary acoustic cues for prosody are pitch, duration and energy. Other acoustic cues like voice quality are useful in general but are hard to be reliably estimated. The raw f0 and RMS energy feature are obtained using Entropy XWAVE, a commercial software well-known for its high accuracy pitch detector. Duration features are obtained using the time-aligned phoneme transcription either generated by hand or by automatic methods.

It is important to normalize the pitch features such that they are less affected by inter-speaker and intra-speaker difference. The noisy $f_0(t)$ returned by the pitch tracker are first filtered by a 3 mixture Gaussian classifier (with the mixture component means restricted to be $1/2$, 1, and 2 times the utterance mean $\bar{f}_0$) to remove the pitch doubling and halving errors, and are then converted using:

$$\hat{f}_0(t) = \log(f_0(t)/\bar{f}_0 + 1). \tag{5}$$

The $\hat{f}_0(t)$ with probability of voicing (output also from XWAVE) smaller than an empirical threshold are removed since they are normally extracted from non-vocalic frames and are not reliable. Linear interpolation is carried out to recover the complete $\hat{f}_0(t)$ contour where the original measures have been previous removed [8]. $\hat{f}_0(t)$ is further normalized by a MLP-based nonlinear transformation function $\psi(\cdot)$ trained to minimize the mean square error between the transformed feature $\tilde{f}_0(t)$ and a teaching signal that indicates the location of the transcribed pitch accents:

$$\tilde{f}_0(t) = \psi(\hat{f}_0(t)). \tag{6}$$

It is shown in our experiment that this nonlinear transformation has considerably reduced the intra-speaker differences, especially the pitch declination effects (the gradual reduction of mean and variance of f_0 toward the end of a prosodic phrase) which is known to hurt the accent prediction. An example illustrating this nonlinear transformation is given in Fig. 1.

A group of five features are computed as our base feature vector $\vec{x_i}$:

1. unnormalized allophone duration,
2. average unnormalized allophone duration over a window of 3 allophones,
3. average energy over a window of 3 allophones,
4. the delta of the 3-phone-average of the phoneme-wise mean $\tilde{f}_0$,
5. the delta of item 4.

These features are similar to those in the previous works [2, 8] and are shown to give the best performance among a set of around 15 features. Unnormalized duration is used instead of normalized duration because we found it gives better performance [10]. Average unnormalized allophone duration over a window of 3 allophones is a feature that encodes pause duration. The longer the pause is, the more it influences the neighboring phonemes. After $\vec{x_i}$ is computed, they were rotated using principle component analysis (PCA) such that they can be better modeled by diagonal covariance Gaussian PDF. The delta of the rotated feature vectors are attached to make up a 10-dimensional feature vector $\vec{y_i}$ for each allophone.

3.2. Syntactical features

Syntactical feature vector in our system includes syntactical information from a window of 5 words centered at current word. The syntactical information extracted from each of these 5 words includes:

1. parts-of-speech,
2. The number of opening parentheses accumulated before this word,
3. The number of closing parentheses right after this word.

A set of 32 POS tags are used, which are the same as those used in the Penn Treebank. Syntactic parsing is carried out automatically by Charniak's syntactic parser [11]. Since "silence" is annotated in our word transcription, we augmented our parts-of-speech set to include a new label "SIL" which is shown to be very useful for boundary prediction. The "pause" and "breath" cues are among those that are most robust for boundary prediction. If they are not annotated in word transcription, they can be inferred from punctuation or automatic silence/breath detection results. Each POS tag is then mapped to a 33 dimensional binary feature vector. The features for the clause opening and closing are integer-valued and are normalized to real numbers after being divided by a constant. Each MLP input vector hence contains $35 \times n$ syntactic features.

4. Experiments and Results

Our first experiment was carried out on the Boston University Radio News Corpus (RNC), one of the largest corpora designed for study of prosody [3]. In this corpus, a majority of paragraphs are annotated with the orthographic transcription, phone alignments, part-of-speech tags and prosodic labels. In our experiment, only intonational phrase boundary versus non-intonational phrase boundary, pitch-accented versus pitch-unaccented are distinguished.

A leave-one-speaker-out strategy is applied to estimate the system performance. Data used in the experiments are extracted from 4 speakers: F1A, F2B, M1B and M2B (where F/M designates female/male speakers). For each experiment, we used data from one speaker for test and the other three for training. F2B were never left-out because it contains the most training data. The statistics of the speakers are listed in Table 1 and the average (weighted by number of words in each speaker) recognition results are listed in Table 2.

Speakers	F2B	F1A	M1B	M2B
# Utterances	164	51	38	33
# Words	14844	3098	3366	2363
# Accents	6345	1382	1500	1061
# IPBs	2744	497	445	409

Table 1: The number of utterances, number of words, number of accents and number of intonational phrase boundaries (IPBs) for the 4 speakers used in our experiment.

	Accent	Boundary	Acc. Bnd. combined
AM only	76.58	68.23	50.06
LM only	82.67	90.09	76.81
Combined	83.91	93.07	78.42

Table 2: The averaged accent, boundary and accent/boundary combined recognition accuracy (%) for acoustic model only (AM only), language model only (LM only) and acoustic model language model combined systems on the leave-one-speaker-out task on the Radio News Corpus.

As shown in Table 2, the acoustic model only (AM only) results are worse than Wightman's results (84% for accent and 71% for boundary). However, our task is more difficult since our training set contains no utterance spoken by the test speaker. On the other hand, since our GMM-based acoustic model is simpler than Wightman's decision tree acoustic model both in the structure and in the dimensionality of input features, worse results are expected. An advantage of our acoustic model is that it provides better generalizability to unseen data as it can better avoid over-training problems than decision trees due to its simplicity. The language model only (LM only) results are very good. Especially, the boundary recognition rate has reached 90% which is 7% better than the chance level 83%. Accent can also be predicted by syntax with an 82.7% accuracy. Combining acoustic model and language model, we achieved accent recognition accuracy of 84.2% and boundary recognition accuracy of 93%, approaching the agreement rate between different human labelers (85-95% for accent, 95-98% for IPB using ToBI) for both accent and boundary recognition.

Our second experiment tests the recognizer trained on RNC on a subset of prosodically labeled switchboard data [12]. This experiment provides us a preliminary measure on how prosody differs across different speech styles. In this experiments, the intermediate phrase boundary and the intonational phrase boundary are grouped as a single boundary class. Results are reported in Table 3.

As shown in Table 3, both accent and boundary recognition accuracies are significantly better than the chance levels. This result indicates that our system can be used for preliminary prosody labeling of switchboard to ease the human labeling efforts.

	Accent	Boundary
chance	68.0	77.9
AM only	74.48	71.19
LM only	78.71	82.61
combined	78.76	82.61

Table 3: The averaged accent and boundary accuracy (%) for acoustic model only (AM only), language model only (LM only) and acoustic model language model combined systems on a subset of switchboard corpus.

5. Conclusions

In this paper, we propose a Maximum Likelihood prosody recognizer consisting of a GMM-based acoustic model that models the distribution of phone-level acoustic-prosodic observations (pitch, duration and energy) and a MLP-based language model that models the stochastic dependence between prosody and syntax. Our experiments on Radio News Corpus have demonstrated the effectiveness of syntax-based language model . The acoustic model alone gives moderate performance but is shown to capture complementary information which is useful to improve the overall system performance. Our prosody recognizer is able to achieved 93% IPB recognition accuracy and 84% pitch accent accuracy in a leave-one-speaker-out task, which are significantly better than previous results and are approaching the agreement rate among different human labelers. The recognizer trained on RNC is tested on a subset of switchboard corpus and achieved accuracies significantly better than the chance levels.

6. References

[1] T. Cho. *Effects of Prosody on Articulation in English*. PhD thesis, UCLA, 2001.

[2] C. Wightman and M. Ostendorf. Automatic labeling of prosodic patterns. *IEEE Trans. Speech and Audio Processing*, 2(4):469–481, Oct. 1994.

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

[4] J. Hirschberg. Pitch accent in context: Predicting intonational prominence from text. *Artificial Intelligence*, 63(1-2), 1993.

[5] S. Arnfield. *Prosody and syntax in corpus based analysis of spoken English*. PhD thesis, University of Leeds, 1994.

[6] K. Ross and M. Ostendorf. Prediction of abstract prosodic labels for speech synthesis. *Computer Speech and Language*, 10:155–185, Oct. 1996.

[7] M. E. Beckman and G. M. Ayers. *Guidelines for ToBI Labelling: the Very Experimental HTML Version*. www.ling.ohio-state.edu/research/phonetics/E_ToBI/singer_tobi.html, 1994.

[8] R. Kompe. *Prosody in Speech Understanding Systems*. Springer-Verlag, 1997.

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

[10] A. Batliner, E. Noth, J. Buckow, R. Huber, V. Warnke, and H. Niemann. Eliminating downstep in prosodic labeling of American English. In *ISCA Workshop on Prosody in Speech Recognition and Understanding*, pages 23–28, 2001.

[11] E. Charniak. A maximum-entropy-inspired parser. In *Proc. of NAACL*, 2000.

[12] S. Chavarria, T. J. Yoon, and J. Cole. Acoustic differentiation of ip and ip boundary levels: Comparison of l- and l-l% in the switchboard corpus. in review.