Prosody-Dependent Acoustic Modeling Using Variable-Parameter Hidden Markov Models

Jui-Ting Huang[1], *Po-Sen Huang*[1], *Yoonsook Mo*[2], *Mark Hasegawa-Johnson*[1], *Jennifer Cole*[2]

[1]Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, USA
[2]Department of Linguistics, University of Illinois at Urbana-Champaign, USA
jhuang29,huang146,ymo,jhasegaw,jscole@illinois.edu

Abstract

As an effort to make prosody useful in spontaneous speech recognition, we adopt a quasi-continuous prosodic annotation and accordingly design a prosody-dependent acoustic model to improve ASR performances. We propose a variable-parameter Hidden Markov Models, modeling the mean vector as a function of the prosody variable through a polynomial regression model. The prosodically-adapted acoustic models are used to re-score the N-best output from a standard ASR, according to the prosody variable assigned by an automatic prosody detector. Experiments on the Buckeye corpus demonstrate the effectiveness of our approach.

Index Terms: Prosody-dependent ASR, variable parameter HMM, re-scoring

1. Introduction

Prosody information improves word recognition accuracies in automatic speech recognition (ASR) [1, 2, 3]. Most of the existing schemes that utilize prosody in ASR were based on symbolic prosodic events, such as the presence/absence of pitch accent [1, 2], the break/non-break of prosodic word [3], and intonational boundary types [2]. However, symbolic events are not the only way to annotate prosody; a quasi-continuous prosodic annotation was introduced in [4] to describe the prosody perception of untrained listeners to spontaneous speech. The prominence score (P-score) or boundary score (B-score), ranging from 0 to 1, is the fraction of listeners who hear the word as prominent or as followed by a boundary. It implies the confidence level of listeners perceiving a prosodic event and can be interpreted as the level of prominence or the degree of disjuncture. Since the continuous variables inherently carry more information than the discrete ones, we aim at finding an adequate way to utilize this alternative prosodic measure to improve ASR.

Speakers modulate the temporal as well as the spectral characteristics of consonants and vowels in their speech to encode prosodic structures. For example, in a series of studies, Mo and her colleagues [4, 5, 6] demonstrated that the perceived prominence significantly correlates with the vowel duration and intensity and the formant measures. Considering the influence of prosody on spectral features, we propose a prosody-dependent acoustic modeling for Gaussian Mixture Hidden Markov Models, in which the model parameters vary as functions of prosodic conditions, hence called Variable-Parameter HMMs (VPHMM) [7]. Specifically, the mean vectors of Gaussian mixture components in each HMM state are modeled as a polynomial function of prominence scores associated with the current utterance.

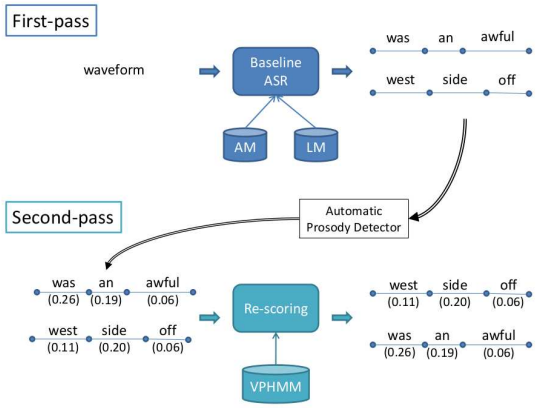

Figure 1: *Diagram of the proposed ASR system.*

Our idea is an extension of Fujinaga's work [8]. In their work, the mean vectors of Gaussian distributions were a linear function of the pitch value at the current time frame, which improved the isolated Japanese word recognition. While they used pitch as a proxy of prosody structure, we go further and use a high-level prosodic variable, P-score, as the auxiliary feature. In our work, the P-scores are assigned automatically to each word by a Support Vector Regression (SVR) method [9], given the suprasegmental features extracted from the word and its context. Since our automatic prosody detector requires the word/phone boundary information for prosodic feature extraction, we employ a two-pass rescoring framework as an improved ASR system (Section 2). The competitive word hypotheses generated from a standard ASR system are re-scored using the prosodically-adapted acoustic models, VPHMM (Section 3), according to the automatically-assigned P-scores (Section 4.2).

2. Two-Pass ASR

Figure 1 illustrates the two-pass ASR scheme in the *recognition* stage. The first pass generates N-best word hypotheses using a standard decoder, and the output includes the information of word and phone boundaries. For each word in the hypothesis, an automatic prosody detector extracts related prosodic features (pitch, intensity and duration) and accordingly predicts a continuous prominence score (P-score), indicated by the number in parenthesis under the word in the figure. The second pass re-evaluates each word in the N-best list using a new acoustic

model (VPHMM) in which parameters are adapted through a polynomial function of the prominence score associated with that word. The N-best list is then re-ranked according to the updated total scores.

During the *training* stage, a SVR-based prosody detector is trained using a relatively small subset of transcribed data (see Section 4.1). Also, the coefficients of the polynomial function for each HMM state in VPHMM are estimated using maximum likelihood criteria. Unlike conventional GM-HMMs, the required observations for VPHMM parameter estimation include not only MFCC/PLP coefficients but also P-scores. Therefore, SVR is also used in the training stage to assign P-scores to words in the training set for the purpose of VPHMM training.

3. Variable-Parameter HMMs

3.1. Formulation

In conventional Gaussian Mixture Hidden Markov Models, each phoneme is modeled as a HMM, and the observations (MFCC or PLP vectors) in each HMM state i are modeled as a multivariate Gaussian mixture distribution with K mixture components,

$$p(x_t|s_t = i) = \sum_{k=1}^{K} w_{ik} b_{ik}(x_t), \tag{1}$$

where $b_{ik}(x_t) = \mathcal{N}(x_t; \mu_{ik}, \Sigma_{ik})$ is the k-th Gaussian distribution with mean μ_{ik} and covariance Σ_{ik}, and w_{ik} is the associated weight.

In our VPHMM, the state output model still follows the form of Gaussian Mixture distribution, except that the mean vector here varies as a function of auxiliary variables. In this paper, the auxiliary feature is the P-score y_t. The value of $y_t \in [0, 1]$ varies with time, determined by a prosody detector. Specifically, y_t varies its value word by word, therefore multiple frames corresponding to the same word in an utterance share the same P-score. Consequently, the state output model becomes a conditional Gaussian Mixture model,

$$p(x_t|s_t = i, y_t) = \sum_{k=1}^{K} w_{ik} b_{ik}(x_t|y_t), \tag{2}$$

where $b_{ik}(x_t|y_t) = \mathcal{N}(x_t; \mu_{ik}(y_t), \Sigma_{ik})$, and μ_{ik} and P-score y_t are related through a N-th order polynomial regression model,

$$\mu_{ik}(y_t) = \sum_{n=0}^{N} \mathbf{c}_{ikn} {y_t}^n, \tag{3}$$

where $\mathbf{c}_{ikn}$ is a $D \times 1$ coefficient vector corresponding to each MFCC/PLP dimension, and y_t is a scalar ranging from 0 to 1.

3.2. Parameter estimation

The parameters are estimated using the maximum likelihood criterion. The Expectation-Maximization (EM) algorithm is adopted similar to conventional GM-HMMs. The major difference is at the state output model; instead of mean vectors, the N-th order polynomial coefficients $\mathbf{c}_{ik} = [\mathbf{c}_{ik0}^T \cdots \mathbf{c}_{ikN}^T]^T$ from Equation (3) are to be estimated. With R training utterances, the auxiliary function for the state output model in the E-step of EM algorithm is defined as:

$$Q_b(\lambda, \lambda') = \sum_{r=1}^{R} \sum_{i=1}^{S} \sum_{k=1}^{K} \sum_{t=1}^{T_r} \gamma_t^r(i,k) \log b_{ik}(x_t^r|y_t^r), \tag{4}$$

where S is the number of different states, K is the number of Gaussian components, T_r is the number of frames in utterance r. $\gamma_t^r(i,k) = p(s_t^r = i, m_t^r = k|X^r, \lambda')$ is the probability of being in state i and mixture component k at time t given the observation sequence X of utterance r, which is calculated using the same manner as in the conventional HMM.

The M-step maximizes $Q_b(\lambda, \lambda')$ in Equation (4) with respect to $\mathbf{c}_{ik}$. Taking derivative of Equation (4) with respect to $\mathbf{c}_{ikn}$ ($n \in \{0 \dots N\}$), we get

$$\sum_{l=0}^{N} A_{ik}(n,l) \cdot \mathbf{c}_{ikl} = \mathbf{b}_{ikn}, \tag{5}$$

where

$$\mathbf{A}_{ik}(n,l) = \sum_{r=1}^{R} \sum_{t=1}^{T_r} \gamma_t^r(i,k) \cdot \Sigma_{ik}^{-1} \cdot (y_t^r)^{(n+l)} \tag{6}$$

and

$$b_{ikn} = \sum_{r=1}^{R} \sum_{t=1}^{T_r} \gamma_t^r(i,k) \cdot \Sigma_{ik}^{-1} \cdot (y_t^r)^n \cdot x_t^r \tag{7}$$

are the accumulating statistics.

Equation (5) for all n needs to be satisfied, and consequently for each state i and mixture k, a linear equation system described in the following matrix compact form needs to be solved:

$$\mathbf{A}_{ik}\mathbf{c}_{ik} = \mathbf{b}_{ik}, \tag{8}$$

where $\mathbf{A}_{ik}$ is a $D(N+1) \times D(N+1)$ matrix in which the $D \times D$ block matrix at the (m+1)-th row and the (n+1)-th column is $\mathbf{A}_{ik}(m,n)$ ($m, n \in \{0 \dots N\}$), and $\mathbf{b}_{ik} = [\mathbf{b}_{ik0}^T \dots \mathbf{b}_{ikN}^T]^T$. Since our HMM system employed diagonal covariance matrices, the polynomial coefficients in different dimensions can be obtained by solving the decoupled linear system for each dimension d ($d = 1 \dots D$),

$$\mathbf{A}_{ik}^{(d)} \mathbf{c}_{ik}^{(d)} = \mathbf{b}_{ik}^{(d)}, \tag{9}$$

where the superscript (d) indicates the d-th diagonal element of each block element in $\mathbf{A}_{ik}$, and the d-th dimension of each element in $\mathbf{c}_{ik}$ or $\mathbf{b}_{ik}$.

4. Experiments

4.1. Corpus and prosody transcription

We perform our experiments on the Buckeye corpus of the spontaneous speech of American English [10]. For ASR experiments, the speech are segmented whenever there is a turn of interviewer or a silence longer than 1 second, based on the given transcription. We randomly choose 36 speakers, from which 80% (around 15 hours) are used as the traning set for acoustic modeling and 1636 utterances sampled from the rest of 20% as a multi-speaker test set. We leave 4 speakers for cross-speaker experiments for future use.

For prosodic transcription, a total of about 8 minutes of speech from each speaker (8 min * 39 speakers = 312 min) was extracted. The extracted speech from each speaker was further divided into 14 to 16 short speech excerpts (around 30 second long per each excerpt). For prosody transcription, 37 monolingual native speakers of American English from undergraduate linguistics courses at the University of Illinois at Urbana-Champaign were recruited. In a series of transcription tasks, listeners were asked to mark the locations of prosodic prominence and boundary on words in a printed transcript where all

punctuation and capitalization were removed. They were only provided with minimal definitions of prosodic prominence and boundary. Then they marked words heard as prominent and words followed by a boundary in real time, while they were listening to speech excerpts in order as said.[1] Each excerpt was transcribed by 15 untrained, non-expert listeners. Transcriptions are pooled across listeners and each word is assigned a probabilistic prominence (P-score) and boundary (B-score). The P- and B-scores range from 0 to 1, depending on the fraction of listeners who marked the word as prominent or followed by a boundary. For example, if no listener heard a word as prominent, then the word is assigned "0" as P-score and if all 15 listeners heard a word as followed by a boundary, then the word is assigned "1" as B-score.

4.2. Automatic prosody detector

As of the time of this paper, the prosodic transcription for 18 out of 39 talkers are finished pre-processing and ready for use. We partition the transcribed data into the training and test set, with a ratio of 4:1, for automatic prosodic labeling. The test set here is merely for the purpose of evaluation of our automatic prosody detection method. To predict the value of P-score by a number of prosodic features, we use Support Vector Regression (SVR) implemented by LIBSVM [11] to learn the functional dependence of P-score on the prosodic features. The predicted values by SVR are then clipped to the feasible region of P-score, [0,1].

Before feature extraction, we obtain the phonetic transcription by forced-aligning speech to the word transcription using a standard dictionary, the same one used for ASR such that the phonetic classes for VPHMM training and decoding are consistent. The prosodic features are then extracted based on the obtained phonetic transcription. The features include the averaged pitch, mean squared energy, and duration, as well as their minimum and maximum value within various kinds of segments such as the stressed vowel, the last vowel, the whole word and the following word. The differences of pitch and energy between the current word and the following word and the pause duration are also considered. Several pre-processing and normalization schemes are applied to features:

- Pitch: fundamental frequency is obtained using ESPS *get_f0* utility with its default values of 10 ms frame shift and 7.5 ms correlation window. A median filter is applied to the logarithm of the estimation ouput for pitch contour smoothing, followed by a local normalization within a window of 400 ms for each frame to remove the intonational phrase effect and the speaker effect, as this work focuses on prominence phenomena.
- Duration- and energy-related features are z-normalized with respect to the phone type, using data pooled from all speakers, to minimize effects due to vowel quality.

SVR shows the squared correlation of 0.48 between P-score and extracted prosodic features on the training set and 0.33 on the test set. Although the correlation is not perfect, we use the same prosodic detector to assign P-scores on both training and N-Best data thus at least the training and test conditions for ASR have a matched prosodic condition. It is worth mentioning that the distribution of the predicted values are quite different from that of the transcribed labels, as shown in Figure 2. The distribution of automatic P-scores appear to concentrate in a lower range (mean = 0.0881) than the transcribed ones (mean = 0.1861).

[1]For more details about the methodology, please refer to [4].

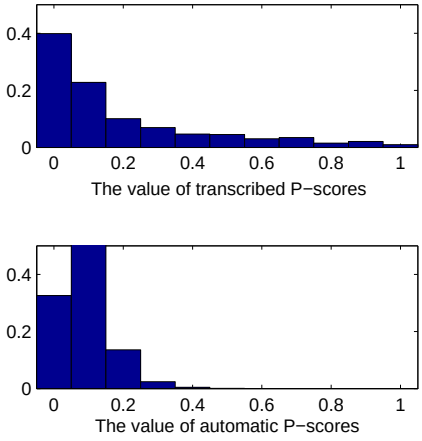

Figure 2: *P-score histograms in the transcription (upper panel) and in the prediction (lower panel).*

4.3. Baseline ASR

We extract 13 Perceptual Linear Prediction (PLP) coefficients, as well as the first-order and second-order derivatives as the front-end spectral features. The frame length is 25 ms and the frame shift is 10ms. Cepstral mean and variance normalization are applied per utterance. We use context-independent models (monophones) for our baseline model as well as VPHMMs. We choose monophones rather than triphones simply because the triphone models would require parameter clustering for polynomial coefficients in VPHMMs, which complicates the implementation of our main idea. A three-state HMM is adopted for each monophone, the silence model and the noise model, and an one-state HMM for the short pause model. The state output models are Gaussian mixture models with 33 components which all emloy diagonal covariance matrices. Since we does not focus on language models, we train a simple language model using both the training and testing transcription. The word recognition accuracy for the baseline is 57.30%.

4.4. VPHMM Results

We generate ten best hypotheses per utterance, and report the rescoring results on the test set with the first- and second- order mean polynomial VPHMMs (corresponding to $N = 1, 2$ in Equation (3)), as shown in Table 1. The two-tailed significance test indicates that the second order polynomial VPHMM is a better system than the baseline at the significant level of 2.9%; the first order polynomial is a better system at the significant level of 5.2%. For reference, the best possible accuracy that can be achieved from the N-Best list (N=10) is 65.59%, which is the upper bound of the improved performance.

Table 1: *Rescoring results with the first- and second-order mean polynomial in VPHMMs.*

	word accuracy ($\%$)	significance level ($\%$)
baseline	57.30	-
1st order	57.79	5.2
2nd order	57.87	2.9

Here we present an example from the rescoring results in which the first-pass ASR errors are corrected by our prosodically-adapted acoustic models. For this particular utterance, the highest ranked hypothesis in the N-Best list of the first-pass ASR is “um was an awful”, whereas the ground truth “um west side off” is the third candidate in the N-best list. The P-Scores predicted by SVR are indicated in the parenthesis after each word:

um (0.17) was (0.26) an (0.19) awful (0.06)
um (0.22) west (0.11) side (0.20) off (0.06)

Even though the prosodic detection is independent of word types, the predicted P-scores appear to be slightly different in the two utterances, even for the same word (e.g. ”um”). This is due to the different time span of phone and word segments, and the normalization of prosodic features by different phone types.

VPHMMs in the second pass adapt the mean vectors of Gaussian components according to the P-scores, thus adjust the acoustic likelihood of each phone. One of the largest adjustments comes from the decrease in the acoustic likelihood for phone /z/ in the word ”was”. This is likely due to the high P-score of phone /z/ giving rise to an adapted model, in which a higher energy distribution in the very low frequency range is expected, inconsistent with the actual spectrogram. As a result, the candidate “um west side off” has the highest total score among all hypotheses, and hence the errors in the first-pass recognition are corrected.

5. Discussion

As rescoring improves the recognition accuracies, VPHMMs appear to capture the variations of spectral realization caused by the word prominence. To our best knowledge, this is the first time that the prosody information at the level higher than lexical stress is utilized for spontaneous speech recognition in English [2]. We would like to point out a few directions which might improve the performance of the present work:

- The automatic prosody detector built for this work is not optimal. The prosodically-adapted acoustic models can be made more reliable by increasing the correlation between the predicted P-score and the true P-score. It would be helpful to search for useful predictors for P-score prediction in addition to the current feature set.
- We expect that modeling the effect of prosody on context-dependent models will provide more improvements than context-independent models (i.e. monophones). However, the parameter clustering problem needs to be solved, as the number of parameters for triphones will be much higher than monophones.
- The current application is restricted to N-best or word lattice re-scoring tasks, which are subject to a limited improvement due to a fixed small search space. It is possible to do one-pass decoding instead. For example, the syllable boundaries can be detected in the first pass, based on which prosodic features are extracted for prosody detection. Then an one-pass ASR with VPHMMs can be directly applied.
- In addition, the prosodic boundaries (B-scores) has not been considered in this work. However, P-scores and B-scores are dependent on each other. Ideally one would like to model the variation of the spectral feature distribution as a function of both two prosodic variables.

[2] Wang et al. [12] incorporated a four-class lexical stress model for ASR on conversational telephone speech.

6. Conclusions

In summary, we proposed variable-parameter HMMs for prosody-dependent acoustic modeling and showed how it can be used to improve ASR performances. We demonstrated the effectiveness of prosody-dependent acoustic modeling on the Buckeye corpus of the spontaneous speech of American English by 0.49-0.57% absolute improvement of word recognition accuracy in a N-best rescoring task. We used one corrected first-pass error as an example to show the contribution of prosody to ASR. We believe that utilizing quasi-continuous prosody annotations will be a promising direction for prosody-dependent ASR.

7. Acknowledgements

We thank J.H. Chu and X. Zhou for helpful discussions. This work was funded by the National Science Foundation Grant CCF 04-26627.

8. References

[1] S. Ananthakrishnan and S. Narayanan, “Unsupervised adaptation of categorical prosody models for prosody labeling and speech recognition,” *IEEE Transactions on Audio, Speech & Language Processing*, vol. 17, no. 1, pp. 138–149, January 2009.

[2] K. Chen, M. Hasegawa-Johnson, A. Cohen, S. Borys, S.-S. Kim, J. Cole, and J.-Y. Choi, “Prosody dependent speech recognition on radio news corpus of american english,” *IEEE Transactions on Audio, Speech & Language Processing*, vol. 14, no. 1, pp. 232–245, 2006.

[3] J.-T. Huang and L. shan Lee, “Prosodic modeling in large vocabulary mandarin speech recognition,” in *The Ninth International Conference on Spoken Language Processing*, 2006, pp. 513–516.

[4] Y. Mo, J. Cole, and E.-K. Lee, “Nave listeners prominence and boundary perception,” in *The proceedings of the fourth Speech Prosody*, 2008.

[5] Y. Mo, “Duration and intensity as perceptual cues for nave listeners prominence and boundary perception,” in *The proceedings of the fourth Speech Prosody*, 2008.

[6] Y. Mo, J. Cole, and M. Hasegawa-Johnson, “Prosodic effects on vowel production: evidence from formant structure,” in *The proceedings of the Tenth Interspeech*, 2009.

[7] X. Cui and Y. Gong, “A study of variable-parameter gaussian mixture hidden markov modeling for noisy speech recognition,” *IEEE Transactions on Audio, Speech & Language Processing*, vol. 15, no. 4, pp. 1366–1376, 2007.

[8] K. Fujinaga, M. Nakai, H. Shimodaira, and S. Sagayama, “Multiple-regression hidden markov model,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2001, pp. 513–516.

[9] A. J. Smola and B. Schölkopf, “A tutorial on support vector regression,” *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, 2004.

[10] M. A. Pitt, K. Johnson, E. Hume, S. Kiesling, and W. Raymond, “Buckeye corpus of conversational speech (2nd release),” 2007. [Online]. Available: www.buckeyecorpus.osu.edu

[11] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001, software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[12] C. Wang and S. Seneff, “Lexical stress modeling for improved speech recognition of spontaneous telephone speech in the jupiter domain,” in *The Seventh European Conference on Speech Communiction Technology*, 2001, pp. 2761–2764.