An Automatic Prosodic Event Detector Using MSD HMMs for Persian Language

Fatemeh Sadat Saleh(✉), Boshra Shams, Hossein Sameti, and Soheil Khorram

Sharif University of Technology, Tehran, Iran
{saleh, bshams, khorram}@ce.sharif.edu,
sameti@sharif.edu

Abstract. Automatic detection of prosodic events in speech such as detecting the boundaries of Accentual Phrases (APs) and Intonational Phrases (IPs) has been an attractive subject in recent years for speech technologists and linguists. Prosodic events are important for spoken language applications such as speech recognition and translation. Also in order to generate natural speech in text to speech synthesizers, the corpus should be tagged with prosodic events. In this paper, we introduce and implement a prosody recognition system that could automatically label prosodic events and their boundaries at the syllable level in Persian language using a Multi-Space Probability Distribution Hidden Markov Model. In order to implement this system we use acoustic features. Experiments show that the detector achieves about 73.5 % accuracy on accentual phrase labeling and 80.08 % accuracy on intonation phrase detection. These accuracies are comparable with automatic labeling results in American English language which has used acoustic features and achieved 73.97 % accuracy in syllable level.

Keywords: Prosody · Accentual · Intonational · Hidden Markov Model · Multi-Space Probability Distribution HMM

1 Introduction

Prosody is comprised of information including variations of the pitch contour, rhythm, and lexical stress patterns of spoken utterance. These prosodic events are suprasegmental effects that operate at a level higher than the local phonetic context. Speakers use these prosodic effects to provide cues to the listener and help for better interpretation of their speech and better spoken language understanding [1].

In this paper, we introduce and build an automatic detector for labeling prosodic events trained with a small number of manually labeled samples, and after that could automatically label large corpora in real time.

Previous models were widely focused on prosody labeling at the phone level. It is important to note that, prosodic characteristics are often not related to the information of the phonetic level and are only associated with phrase semantics [2]. Therefore, it is more preferable to consider prosodic features at syllable level.

Wightman and Ostendorf [3] proposed Automatic detection of prosodic events for the first time at syllable level. Their modeling was based on posterior probabilities

A. Movaghar et al. (Eds.): AISP 2013, CCIS 427, pp. 234–240, 2014.
DOI: 10.1007/978-3-319-10849-0_24

computed from acoustic features using decision tree. In [4] the prosodic features were combined and applied to time series modeling framework and in works such as [5], GMM and ANN have been used to classify phrase boundaries using features extracted from syllable level. In [6] a prosody recognition system that detects prosodic boundaries at word and syllable level in American English language using CHMM was introduced. It is believed that the acoustic features of prosodic events consist of multiple streams of information that are correlated but are not always synchronous [6]. Thus, we can take advantage of HMM models with multiple feature streams to deal with asynchrony between acoustic features.

Moreover, one of the problems with fundamental frequency ($F0$) modeling as one of the most important prosodic features is that $F0$ values are not defined in the unvoiced region. MSD-HMM [7] is a practical solution to deal with this problem. In this paper we design an automatic prosodic labeling that can detect Intonational Phrases and Accentual Phrases boundaries in syllable level for Persian language, based on MSD-HMM modeling.

The remainder of this paper is organized as follows. The prosodic event definitions are proposed in Sect. 2. Section 3 presents the proposed methodology. Section 4 presents some experimental results, and finally some conclusions are given in Sect. 5.

2 Prosodic Events Definitions

There are two prosodic constituents in the prosodic structure of the Persian language. The smallest prosodic unit in Persian is the Accentual Phrase (APs) [8]. According to linguistic experts, 6 types of APs are defined for Persian language. Although there may be minor differences in the definitions of these types, in this research we define APs as shown in Table 1. The next level of prosodic events in Persian language is Intonational Phrase (APs) that consists of one or more APs.

Table 1. Persian prosodic AP's definitions [8]

H*	For initially stressed words and phrases and also one-syllable words
L+H*	For words and phrases with final stress
l,h	The part of an Accentual Phrase between the pitch accent and the AP end is handled by a boundary tone, which can be high or low, named here as h and l
L+	For words and phrases losing their stress in sentence
p	For showing silence in sentences

The right boundary of an IPs is determined with a low or high boundary tone that is specified with L% or H% [8].

3 Methodology

In this part, we explain our methodology in two sections. First, we introduce the feature used and then the structure of the model.

3.1 Features

In this paper three features are used for modeling. Each feature is introduced in below. In order to compute feature vectors, the frame length is considered to be 5 ms.

Fundamental Frequency (*F*0). Fundamental Frequency is one of the most important characteristics of the speech signal. In voiced frames, speech signal is periodic with period T in time domain. The numeric value of $F0$ in each frame is $F_0 = \frac{1}{T}$. $F0$ with its first and second derivative are applied as features in three different streams in HMM model. These features efficiently affect detector performance. The speech analysis software used throughout this work to extract pitch track is Praat (Boersma and Weenink 2007) [9].

$F0$ values are not defined in the unvoiced region and the observation sequence of an $F0$ pattern is composed of one-dimensional continuous value and discrete value which represent "unvoiced". However, we cannot apply both the conventional discrete and continuous HMMs to such observations like $F0$. In this paper we model the fundamental frequency ($F0$) with a kind of HMM in which the state output probabilities are defined by Multi-Space Probability Distributions. This kind of HMM model is called MSD-HMM [7].

MSD-HMM can model the sequence of observation vectors with different dimensionalities. So we can model observed $F0$ values in a one-dimensional space and the "unvoiced" segments in a zero-dimensional space.

Intensity. Intensity is a measure of the energy flux, averaged over the period of the signal. To obtain the intensity values we have employed these steps: first hamming windows of the length 400 samples are applied to the signal and the result is squared. Then an averaging is performed on the results. Consecutive frames are processed with a shift of 80 samples.

Duration. In each syllable, the repetition of vowel phonemes determine the duration of that syllable. Therefore, we can generate a vector of duration values for each utterance.

3.2 MSD-HMM

In order to model $F0$ patterns, we assume two spaces, named Ω_1 and Ω_2 representing one-dimensional space for $F0$ in voiced regions and single zero-dimensional space for the unvoiced regions, respectively. In general, we consider Ω_g as one of G sub-spaces with its own dimensionality and probability w_g of observation space Ω of an event E, so we have,

$$\Omega = \bigcup_{g=1}^{G} \Omega_g, \quad \sum P(\Omega_g) = 1 \tag{1}$$

If the dimensionality of each space n_g, is greater than zero, then it has a probability distribution function $N_g(x)$ and if it equals zero, this sub-space contains only one sample point. Here, event E is representing an observation O that consists of a set of space indices X and a random variable $x \in R^n$, that is $O = (X, x)$.

Note that dimensionality of each sub-spaces is depicted by n in X and observation probability of o is defined as below,

$$b(o) = \sum\nolimits_{g \in s(o)} w_g N_g(V(o)) \tag{2}$$

where $s(o) = X$, $V(o) = x$ and $N_g(V(o))$ is a pdf in the sub-space in which random variable x is distributed. In Eq. (2), w_g is the mixture weight for g-th Gaussian, so it is considered as the prior probability of g-th sub-space.

As described before, in order to obtain an automatic prosody recognizer, $F0$ contour curve plays an important role in recognition process, so since the observation sequences of $F0$ patterns are regarded as one-dimension continuous values in voiced regions and as a discrete symbol in unvoiced regions, we employ multi-space distribution probability HMM (MSD-HMM) for modeling discontinuity of fundamental frequency.

To start modeling, first we consider an HMM model λ of the appointed type with N number of states in which the initial states, transition and the state output probability distribution is specified as, $\pi = \{\pi_j\}_{j=1}^{N}$, $A = \{a_{ij}\}_{i,j=1}^{N}$ and $B = \{b_i(.)\}_{i=1}^{N}$ respectively, where,

$$b_i(o) = \sum\nolimits_{g \in s(o)} w_{ig} N_{ig}(V(o)),\ i = 1, 2, \ldots, N \tag{3}$$

Each state i has G pdfs $N_{i1}(.), N_{i2}(.), \ldots, N_{iG}(.)$ with their corresponding weights $w_{i1}(.), w_{i2}(.), \ldots, w_{iG}(.)$, so that we can calculate the observation probability of sequence O as follows;

$$P(o|\lambda) = \sum_{allq} \prod\nolimits_{t=1}^{T} a_{q_{t-1}q_t} b_{q_t}(o_t) = \sum_{allq,X} \prod\nolimits_{t=1}^{T} a_{q_{t-1}q_t} w_{q_t X_t} N_{q_t X_t}(V(o_t)) \tag{4}$$

where $q = \{q_1, q_2, \ldots, q_T\}$ denotes state sequence, $X = \{X_1, X_2, \ldots, X_T\} \in \{s(o_1), s(o_2), \ldots, s(o_T)\}$ is a sequence of possible space indices for the observation sequence O and $a_{q_0 j} = \pi_j$. The modified forward and backward algorithms are utilized for computing observation probabilities. The modified Viterbi algorithm is utilized for the decoding. These algorithms are similar to traditional HMM algorithms [7].

3.3 Structure of Model

The main structure of the proposed automatic prosody detector system is similar to an automatic speech recognizer. In fact the prosodic event sequences with maximum likelihood probability are defined as below,

$$P^* = \underset{P}{\operatorname{argmax}}\, p(A|P) \cdot p(P) \tag{5}$$

where $A = \{a_1, a_2, \ldots, a_n\}$ is set of acoustic features and P is a candidate sequence of prosodic events. Also $p(P)$ is the probability of a given prosodic event sequences.

Finally a prosodic event sequences $P^* = \{p_1^*, p_2^*, \ldots, p_n^*\}$ with the maximum probability will be found through the Viterbi algorithm. In this paper, we have used a left to right MSD-HMM with 5 states and 5 streams. Three streams are $f0, \Delta f0, \Delta^2 f0$ each consisting of two Gaussian distributions. One is defined with zero variance and mean that is for modeling MSD-HMM and the other is a usual Gaussian distribution. Another stream for intensity is a 3-dimensional Gaussian distribution and includes $I, \Delta I, \Delta^2 I$. Finally the fifth stream is a one dimensional Gaussian distribution and contains the duration values. Figure 1 shows the structure of MSD-HMM which used in this paper. As the figure clearly shows, this model is a 5-state left to right HMM with 5 streams.

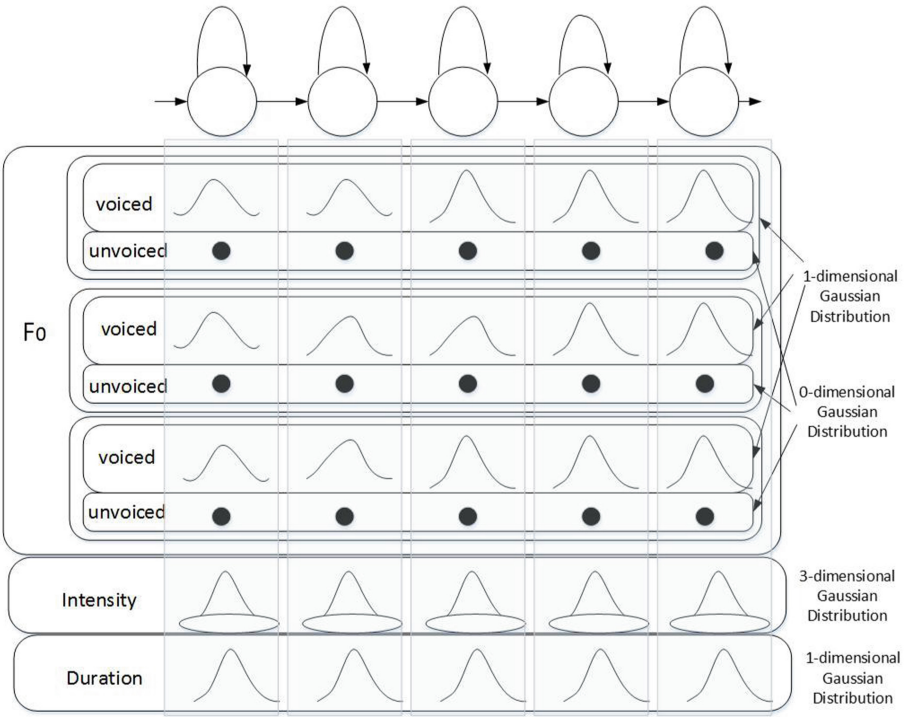

Fig. 1. The structure of used MSD-HMM

4 Experiments

4.1 Experiments Conditions

In order to evaluate the performance of this automatic prosodic detector, a speech database in Persian language was used. This database has been designed after Arctic database in English and contains the speech utterances of two speakers. These utterances are designed in order to satisfy the following conditions [10]:

- Each sentence is short enough to be recorded easily.
- The utterances are phonetically balanced and contain Persian diphones and syllables.

The sentences have been selected from Peykare corpus [11]. The selected sentences contain 5 to 20 words. Final sentences are selected in a way to cover most frequent Persian words and also most frequent syllables. There are some questions and exclamatory sentences, too.

4.2 Experiments Results

Among 1300 available utterances, 1000 utterances were used for training the models and 300 utterances were used for evaluation. There is no overlap between training data and evaluation data. It should be noted that this database is tagged manually with prosodic labels by a linguist. Table 2 shows the results of this automatic prosody detector. The accuracy is the ratio of truly labeled prosodic events by the proposed system to all prosodic events of the utterances.

Figure 2 shows an utterance tagged by prosodic labels. As it is obvious from the figure, pitch contour is an important factor to recognize the prosodic labels. Also the results in Table 2 show that *F*0 is the most important factor in this detector system, but intensity and duration have certain effects on detecting these labels.

In [6] a prosody recognition system that detects stress and prosodic boundaries at the word and syllable level in American English using a coupled Hidden Markov

Table 2. The accuracy results of automatic prosodic event detector

Features	AP	IP
F0	56.5 %	68.8 %
F0, intensity	63.33 %	77.35 %
F0, intensity, duration	73.5 %	80.08 %

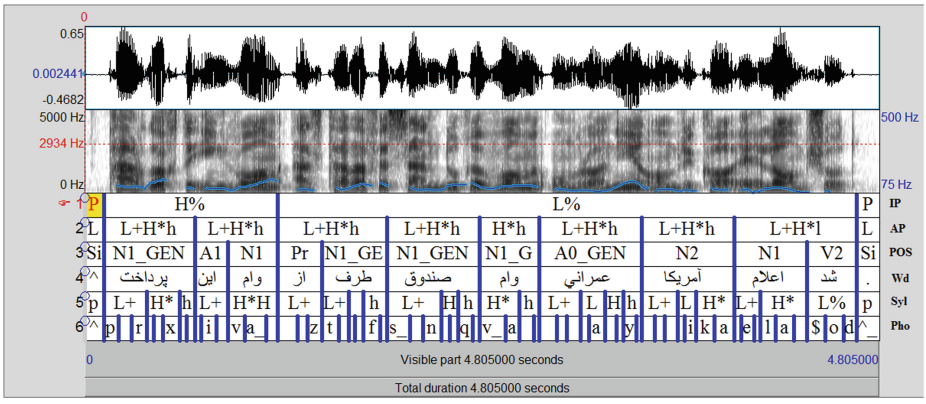

Fig. 2. An utterance which has been tagged with prosodic events.

Model (CHMM) has been introduced. Experiments show that the recognizer achieves 72.03 % accuracy at word level 73.97 % accuracy at the syllable level using acoustic features. The results of our algorithm for Persian language is comparable with these results in American English language.

5 Conclusion

In this paper, we introduced and implemented an automatic prosodic event detector for Persian language. This system is able to detect the Accentual Phrases and Intonational Phrases for Persian utterances. We used fundamental frequency (*F*0), intensity and duration as features. In this article MSD-HMM is used and finally the most probable prosodic sequence is obtained using maximum likelihood criterion. The proposed method reaches 73.5 % accuracy for Accentual Phrase detection and 80.08 % for Intonational Phrase detection.

References

1. Sridhar, V., Bangalore, S., Narayanan, S.: Exploiting acoustic and syntactic features for automatic prosody labeling in a maximum entropy framework. IEEE Trans. Audio Speech Lang Process. **16**(4), 797–811 (2008)
2. Milone, D.H., Rubio, A.J.: Prosodic and accentual information for automatic speech recognition. IEEE Trans. Speech Audio Process. **11**(4), 321–333 (2003)
3. Wightman, C., Ostendorf, M.: Automatic labeling of prosodic patterns. IEEE Trans. Speech Audio Process. **2**(4), 469–481 (1994)
4. Chen, K., Hasegawa-Johnson, M., Borys, S.: Prosody dependent speech recognition with explicit duration modeling at intonational phrase boundaries. In: Proceedings of Eurospeech (2003)
5. Chen, K., Hasegawa-Johnson, M., Cohen, A.: An automatic prosody labeling system using ANN based syntactic-prosodic model and GMM-based acoustic-prosodic model. In: International Conference on Acoustics, Speech and Signal Processing, vol. 1, pp. 509–512 (2004)
6. Ananthakrishnan, S., Narayanan, S.S.: An Automatic prosody recognizer using a coupled multi-stream acoustic model and a syntactic-prosodic language model. In: Proceedings of ICASSP, pp. 269–272 (2005)
7. Tokuda, K., Masuko, T., Miyazaki, N., Kobayashi, T.: Multi-space probability distribution HMM. Inst. Electron. Inform. Commun. Eng. (IEICE) Trans. Inform. Syst. **85**(3), 455–464 (2002)
8. Sadat Tehrani, N.: The Intonational Grammar of Persian. Ph.D. thesis, The University of Manitoba (2007)
9. Praat: doing phonetics by computer, http://www.fon.hum.uva.nl/praat/
10. Bahaadini, S., Sameti, H., Khorram, S.: Implementation and evaluation of statistical parametric speech synthesis for Persian language. In: Proceedings of Machine Learning Signal Processing (MLSP) (2011)
11. Bijankhan, M., Sheikhzadegan, J., Bahrani, M., Ghayoomi, M.: Lessons from creation of Persian written corpus: Peykare. Lang. Resour. Eval. J. **45**(2), 143–164 (2010). Springer, Netherlands