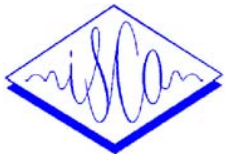

Speech Prosody 2004
Nara, Japan
March 23-26, 2004

Speaker-Independent Automatic Detection of Pitch Accent

Yuexi Ren[1], Sung-Suk Kim[2], Mark Hasegawa-Johnson[1] and Jennifer Cole[3]

[1]Department of Electrical & Computer Engineering
University of Illinois at Urbana-Champaign
{tzhang1, hasegawa}@ifp.uiuc.edu
[2]School of Computer & Information
Yong-In University
sskim@yongin.ac.kr
[3]Department of Statistics
University of Illinois at Urbana-Champaign
Jscole@uiuc.edu

10.21437/SpeechProsody.2004-120

Abstract

This paper presents a novel approach to the automatic detection of pitch accent in spoken English. The approach that we propose is based on a time-delay recursive neural network (TDRNN), which takes into account contextual information in two ways: (1) a delayed version of prosodic and spectral features serve as inputs which represent an explicit trajectory along time; and (2) recursions from the output layer and some hidden layers provide the contextual labeling information that reflects characteristics of pitch accentuation in spoken English. We apply the TDRNN to pitch accent detection in two forms. In the normal TDRNN, all of the prosodic and spectral features are used as an entire set in a single TDRNN. In the distributed TDRNN, the network consists of several TDRNNs each treating each prosodic feature as a single input. In addition, we propose a feature called spectral balance-based cepstral coefficient (SBCC) to capture the spectral characteristic of pitch accentuation. We used the Boston Radio News Corpus (BRNC) to conduct experiments on the speaker-independent detection of pitch accent. The experimental results showed that the automatic labels of pitch accent exhibited an average of 83.6% agreement with the hand labels.

1. Introduction

Pitch accent is a signaling of semantic salience via extruded pitch that stands out from its context, e.g., being high if its neighbors are low, or being low if the neighbors are high. According to a study of pitch movement along syllables, the ToBI standard defines six types of pitch accents: H* (locally highest), L* (locally lowest), L*+H (locally lowest followed by a relatively sharp rise), L*+!H (locally lowest followed by a flat rise), L+H* (a relatively sharp rise from L* followed by a H*) and !H* (a step down onto a pitch accent from a high pitch) [1].

The detection of pitch accent is very important for automatically interpreting spoken language, and thus very useful for the design of spoken dialogue system. Accented words refer to those words on which the pitch accent usually falls in a sentence. Accented words are usually the bearers of important information, especially new information in comparison with the given information. Moreover, unlike other automatic comprehension strategies, such as semantic parsing [2] and key words/phrases spotting [3], the accented word-based comprehension does not depend on the predictability of users' utterances. As a result, the scenario of a spoken dialogue system can be designed to permit content increments by detecting new information using accented words. For example in an intelligent tutoring system, when the computer puts forward the question "where did you see gears before?", if the accented word detected from the response of a child user is "machine", then the computer will check to see if "machine" is in the hypothesized response. If not, then "machine" will be added to the response hypothesis set for the prediction of future responses to the same question. In this way, the accented word detection provides a path for the dialogue system's automatic learning to become smarter and smarter.

In the previous study of automatic labeling of pitch accent, some researchers used the duration, pitch and energy of syllables as the prosodic attributes to train a hidden Markov model whose states represented the labels. The speaker-dependent labeling of BRNC yielded 84% correct detection vs. 13% false detection [4]. Some researchers used pitch and energy on the frame level to train a TDRNN, and the speaker-dependent labeling of BRNC was 91.9% in accuracy for pitch accent and 91.0% for pitch non-accent [5].

2. Proposed approach

2.1. Feature extraction

In addition to the prosodic features including pitch, energy and duration that haven been previously investigated, we also use spectral balance to capture the spectral characteristics of pitch accent. Spectral balance is defined as the intensity increase at higher frequencies (≥ 500 Hz) of vocal speech. The perception of Sluijter, et al. [6] showed that spectral balance was a reliable indicator of stress. If a speaker produced stressed syllables, then the intensity of signals at higher frequencies increased more than the intensity of signals at lower frequencies. The intensity level manipulation of signals at higher frequencies provided stronger stress than the manipulation on the entire frequency band. In spoken English, pitch accent usually occurs on the lexically stressed syllables that are produced with greater vocal effort. Therefore, similar to duration and intensity, spectral balance can be used as a spectral attribute for pitch accent detection.

2.1.1. Spectral balance-based cepstral coefficients

In order to apply spectral balance to our pitch accent detection, we extract features called spectral balance-based

cepstral coefficients (SBCC). Sampled at 11k Hz, the speech signal is pre-emphasized and grouped into frames of 330 samples with a window shift of 110 samples. Multiresolution decomposition is then applied to the speech samples within each frame. The filter bank consists of 14 band-pass filters spanning from 547 Hz to 5000 Hz.

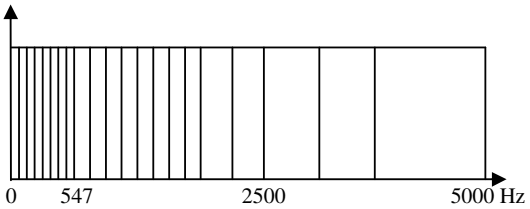

Figure 1: *The multiresolution decomposition bank*.

After decomposing the time-domain speech signals into 14 bands with Daubechies-4 wavelet filter coefficients [7], we compute the signal intensity in each band by:

$$e_m = \frac{1}{N_{\phi_m}} \sum_{n \in \phi_m} y_m^2[n], \qquad 1 \leqslant m \leqslant M \tag{1}$$

where $y_m[n]$ is speech sample n in band B_m, e_m is the intensity of band B_m, M is the total number of bands (M=14), N_{ϕ_m} is the total number of speech samples in set Φ_m, Φ_m is a subset of B_m and

$$\phi_m = \{n \mid |y_m[n]| \geq \frac{1}{2} T_m\} \tag{2}$$

with

$$T_m = \max_{y_m[n] \in B_m} \{y_m[n]\} \cdot \tag{3}$$

Discrete cosine transformation (DCT) is then applied to the intensity of bands to derive the SBCC E_l by

$$E_l = \sum_{m=1}^{M} \log(e_m) \cos[\frac{l(m-0.5)\pi}{M}], \quad 1 \leqslant l \leqslant L \tag{4}$$

where L is the desired length of the cepstrum. The cepstral coefficients of each frame are further subtracted by the mean value over frames to compensate for the disturbance caused by the transmission channel.

2.1.2. Syllable-level features

In our study, each frame is characterized by 13 SBCCs, pitch and log-scaled intensity. Pitch is extracted using the "formant" program in Entropic XWAVES with a probability of voicing (PV) that serves as a confidence measure. The errors caused by pitch doubling and halving are eliminated by deleting those pitch values which fall into the doubling and halving clusters of a 3-mixture Gaussian model. The means of the component Gaussian models are restricted to 1/2μ, μ and 2μ, where μ is the estimated mean value of pitch over the utterance. The overall intensity is normalized by the peak value in order to compensate for the differences in the sound volume across speakers. Then the feature vectors of the frames in a syllable are averaged in a special way to obtain the feature vector of that syllable. The averaging scheme is given by:

$$D_m = \frac{1}{N_{\psi_m}} \sum_{f \in \psi_m} F_m[f] \tag{5}$$

where D_m is the feature vector for syllable S_m, $F_m[f]$ is the feature vector of frame f in S_m, ψ_m is a subset of frames in S_m, N_{ψ_m} is the total number of frames in ψ_m, and

$$\psi_m = \{f \mid |F_m[f]| \geq \frac{1}{2} T_m\}, \tag{6}$$

$$T_m = \max_{F_m[f] \in S_m} \{F_m[f]\} \cdot \tag{7}$$

In addition, the syllable duration is added to complete the feature set. The duration of each syllable is normalized by the number of phonemes in that syllable.

2.2. Normal TDRNN

The detection of pitch accent is modeled by way of TDRNN. TDRNN is a neural network that uses the delayed input to capture the dependence of human perception on the spectra change and the dynamics of speech signals. TDRNN also uses the recurrent circuit(s) to capture the long-term or short-term context information. The context information is very important for pitch accent labeling, because an English word usually has only one primary accented syllable. Therefore, the labeling of a syllable affects and also is affected by the labeling of its adjacent syllables. As shown in Figure 2, TDRNN is a 4-layer back-propagation network with two recursive context layers *a* and *b*, which feed back delayed values from the pitch layer and output layer, respectively. We improve our previous TDRNN design [5] by adding a recursive layer from the output to allow more context information to be captured. The non-recursive input layer *c* processes current and delayed values of the input samples. In the TDRNN, the input layer and the two recursive layers are labeled as level 1, and the other three layers are labeled as levels 2, 3, and 4 along the direction of the arrows. Each node is a weighted sum of the outputs of the nodes on its previous layer(s) through

$$x_{j,h}(t_n) = f\left(\sum_{i \in N_{h-1}} \sum_{k=1}^{K_{i,h-1}} \varpi_{ijk,h-1} \times x_{i,h-1}(t_n - \tau_{ijk,h-1})\right), \tag{8}$$

where $x_{j,h}(t_n)$ is the activation level of node j on layer h at syllable t_n (the syllables are indexed in the chronological order in the utterance), $x_{i,h-1}(t_n - \tau_{ijk,h-1})$ is the activation level of node i on layer h-1 at syllable $t_n - \tau_{ijk,h-1}$, N_{h-1} represents the total number of nodes on layer h-1, $K_{i,h-1}$ represents the total number of time delays for $x_{i,h-1}$, $\omega_{ijk,h-1}$ represents the weight of connection between $x_{j,h}$ and $x_{i,h-1}$ at syllable $t_n - \tau_{ijk,h-1}$, and $f(\bullet)$ is a sigmoid function defined as $f(x) = \frac{1}{1+e^{-x}}$. In this study, we fix the delay $\tau_{ijk,h-1}$ and update the interconnection weight $\omega_{ijk,h-1}$ using the error back-propagation learning algorithm [8]. The difference between the output layer and the pitch layer is compared with a preset threshold. Our algorithm chooses a pitch accent when the difference is larger than the threshold and pitch non-accent otherwise.

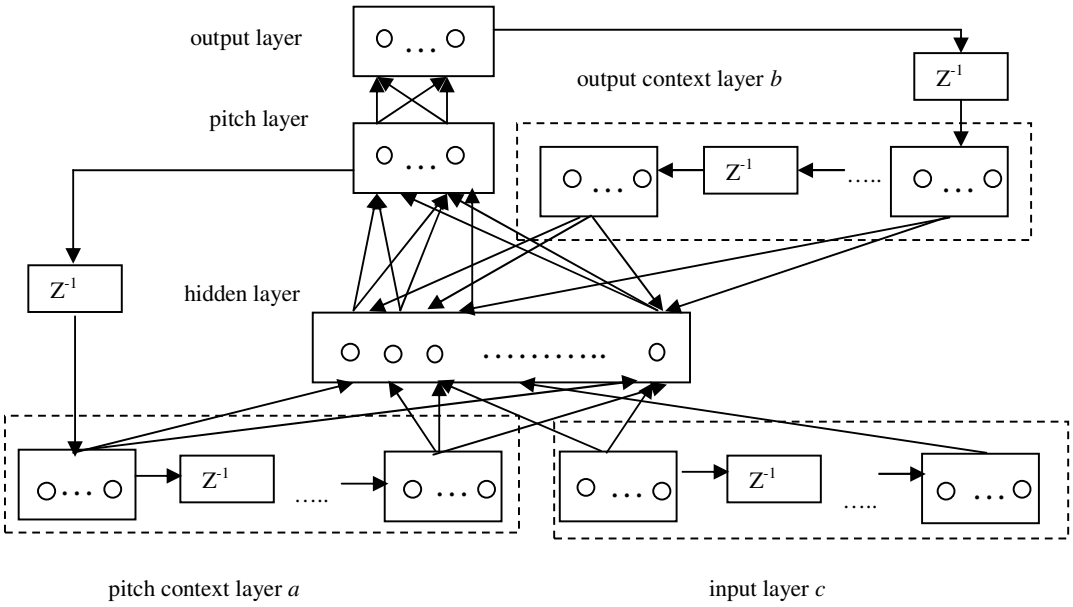

Figure 2: *The structure of TDRNN pitch accent detector (Z^{-1} denotes one syllable time delay).*

The normal TDRNN uses all the features together in a single unified input. That is, the feature vector of a syllable is 16 dimensional consisting of 13 SBCC, pitch, duration and overall intensity.

2.3. Distributed TDRNN

Usually the degree of contribution varies for different features. As listed in Table 1, our experimental results showed that the classifications based on different features have different performance, and the pitch-based classification achieved a higher accuracy than the classifications based on the other features. However, the normal TDRNN does not reveal the difference existing in the contributions of different features to the pitch accent detection. In addition, all of the features used in the normal TDRNN have to use the same structural parameters such as time delays and the number of nodes in a layer. However, our experimental results showed that to achieve a higher accuracy, different structural parameters needed to be adopted for different features. For example, the pitch-based TDRNN used 5 nodes while the duration-based TDRNN used 10 nodes for the hidden layer.

Rather than combining all the features together as an input, the distributed TDRNN consists of four TDRNNs each taking a single feature as the input, and integrates the individual TDRNNs by a 5-layer neural network. In the distributed TDRNN, the outputs of individual TDRNNs form the 4th layer, and the final plant output (the 5th layer) is a weighted sum of the nodes on the 4th layer. Figure 3 depicts the architecture of the distributed TDRNN. The distributed TDRNN offers potential advantages over the normal TDRNN. First, the distributed TDRNN allows the TDRNNs of individual features to adopt different structural parameters (such as number of delays, number of nodes in a layer, decision threshold, etc.) so that each feature alone can achieve a relatively high contribution to the detection performance. Second, before model training, we set the initial weights connecting the nodes on the 4th layer and the node of the 5th layer quite different values to acknowledge the discrepancy in contributions from different features. For example, the interconnection weight for the pitch TDRNN is much larger than the interconnection weights for the TDRNNs of the other features. Then the automatic learning algorithm refines the weights to optimize the contribution assignment of the individual features.

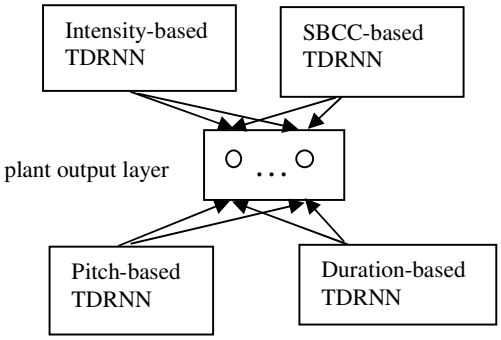

Figure 3: *Structure of distributed TDRNN pitch accent detector.*

3. Experiments

The TDRNN was trained and tested for the purpose of speaker-independent, gender-dependent pitch event recognition using the Boston Radio News Corpus (BRNC). BRNC is a series of

radio stories read by seven professional radio announcers [9], and partially annotated using the ToBI (tones and break indices) prosodic annotation system [1]. We only used female speakers for experiments due to the data insufficiency of male speakers. In addition, because the examples of the pitch accents L*+H and L*+!H in BRNC are sparse, we only used the other four pitch accents for study. We also statistically computed the classification accuracy of pitch non-accent as a measure of the false detection of pitch accent.

The duration of syllables was automatically derived with the help of the toolkit Perl [10]. Because the dictionary in the data package often defines a word with several pronunciations, we first retrieved the corresponding syllables for each word in an utterance according to the dictionary ".prn" file and the word labeling ".wrd" files. The obtained information on word-syllable pairs was further compared with phoneme labeling (the "lbl" and "lba" files) to search for syllable boundaries. We found that some of the word labeling provided by the data corpus was not consistent with the phoneme labeling. The discrepancy of the word labeling and phoneme labeling was manually checked and corrected. We also used Perl to automatically annotate a syllable with pitch accent or non-accent by looking into the tone labeling ".ton" files.

The data corpus includes speech from 3 female speakers, totaling 208 minutes. We used clean speech that makes up to 90% of the entire data corpus. We used approximately 78% of the clean speech for training, and used the other 22% for testing. The data set consists of 8160 pitch accented syllables and 24208 pitch non-accented syllables. The experimental results are summarized in Table 1. The TDRNN based on a feature means that the feature is used as the single input. The test results show that pitch is the most efficient feature, and the distributed TDRNN is more efficient than the normal TDRNN for pitch accent detection.

Table 1: *The classification accuracy of pitch accent and non-accent using the prosodic and spectral features both alone and in combination.*

	Average (%)	Accent (%)	Non-accent (%)
Pitch-based TDRNN	79.79	69.94	84.23
Energy-based TDRNN	73.40	75.92	72.26
Spectral balance-based TDRNN	58.71	57.64	59.19
Duration-based DRNN	53.51	85.18	39.21
Normal TDRNN	81.21	68.27	87.05
Distributed TDRNN	83.64	78.20	86.09

4. Conclusions

This paper presented a novel approach for the automatic detection of pitch accent in spoken English. The detection of pitch accent is significant for both speech recognition and language understanding since the pitch accented words usually contain new and important information. The approach that we proposed was based on TDRNN that combined the prosodic and spectral inputs and the context labeling information. We investigated both the normal TDRNN and the distributed TDRNN to pitch accent detection. The features used for pitch accent detection were pitch, intensity, duration and SBCC that were derived on the syllable level. We used BRNC as the data corpus for experiments on the gender-dependent and speaker-independent pitch accent detection. The experimental results showed that pitch was the most efficient feature for pitch accent detection, and the distributed TDRNN outperformed the normal TDRNN in the classification performance. The accuracy reached as high as 78.2% for accent detection and 86.1% for non-accent detection.

5. Acknowledgement

This work is supported by NSF grant number 0085980. Statements in this paper reflect the opinions and conclusions of the authors, and are not endorsed by the NSF.

6. References

[1] Beckman, M. E., and Ayers, G. M. 1994. *Guidelines for ToBI Labeling*. http://www.ling.ohiostate.edu/phonetics/ToBI/main.html

[2] Seneff, S. 1992. Tina: a natural language system for spoken language applications. *Computational Linguistics* 18(1), 61-86.

[3] Kawahara, T., Lee, C.-H. and Juang, B.-H. 1998. Flexible speech understanding based on combined key-phrase detection and verification. *IEEE Trans. on Speech and Audio Processing* 6(6), 558-568.

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

[5] Kim, S.-S., Hasegawa-Johnson, M., and Chen, K. 2003. Automatic recognition of pitch movements using time-delay recursive neural network. Submitted to *Speech Communication*.

[6] Sluijter, A., van Heuven, V. J., and Pacilly, J. 1997. Spectral balance as a cue in the perception of linguistic stress. *The Journal of the Acoustical Society of America* 101(1), 503-513.

[7] Daubechies, I. 1990. The wavelet transform, time-frtequency localization and signal analysis. *IEEE Trans. on Information Theory* 36(5), 961-1005.

[8] Rumelhart, D. E., McClelland, J. L., and the PDP Research Group. 1986. Learning representations by back-propagation errors. In *Parallel Distributed Processing*. Cambridge, MA: MIT Press, 1: 318-362.

[9] Ostendorf, M., Price, P. J., and Shattuck-Hufnagel, S. 1995. The Boston University radio news corpus. *Lingusitic Data Consortium*.

[10] Wall, L., Christiansen, T., and Orwant, J. 2000. *Programming Perl*. Sebastopol, CA: O'Reilly & Associates.