Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language 43 (2017) 18–33

www.elsevier.com/locate/csl

Intonation classification for L2 English speech using multi-distribution deep neural networks

Kun Li*, Xixin Wu, Helen Meng

The Chinese University of Hong Kong, Hong Kong SAR, China

Received 18 November 2015; received in revised form 11 October 2016; accepted 18 November 2016
Available online 06 December 2016

Abstract

This paper investigates the use of multi-distribution deep neural networks (MD-DNNs) for automatic intonation classification in second-language (L2) English speech. If a classified intonation is different from the target one, we consider that mispronunciation is detected and appropriate diagnostic feedback can be provided thereafter. To transcribe speech data for intonation classification, we propose the RULF labels which are used to transcribe an intonation as *rising, upper, lower* or *falling*. These four types of labels can be further merged into two groups – *rising* and *falling*. Based on the annotated data from 100 Mandarin and 100 Cantonese learners, we develop an intonation classifier, which considers only 8 frames (i.e., 80 ms) of pitch value prior to the end of the pitch contour over an intonational phrase (IP). This classifier determines the intonation of L2 English speech as either rising or falling with an accuracy of 93.0%.

Keywords: Language learning; L2 English speech; Suprasegmental; Intonation; Pitch accent; Deep neural networks

1. Introduction

This work aims to develop speech technologies that assist second language acquisition of English by adult Chinese learners, focusing specifically on suprasegmental phonology (i.e., prosody). English is the lingua franca of our world. It is of prime importance for non-native speakers to acquire communicative competence in English. However, the process of second language acquisition (L2) is interfered by well-established perceptions of sounds and articulations in the primary language (L1). Chinese and English have stark contrasts linguistically. We often observe notable L1 (i.e., Chinese) interferences with L2 (i.e., English) speech in phonetics (i.e., segmental phonology) as well as prosodics (i.e., suprasegmental phonology). While both impede the intelligibility of L2 speech, perceptual studies suggest that suprasegmentals may have a stronger effect (Anderson-Hsieh et al., 1992). The interferences are ingrained with age and hamper acquisition of proficiency, especially for adult L2 learners. Improvements require persistent and individualized perceptual and productive training.

Recent advancements in speech technologies have opened up new possibilities in computer-aided language learning (Eskenazi, 2009). Major thrusts lie in applying automatic speech recognition to the learner's non-native speech

* Corresponding author.
E-mail address: kli@se.cuhk.edu.hk (K. Li), wuxx@se.cuhk.edu.hk (X. Wu), hmmeng@se.cuhk.edu.hk (H. Meng).

http://dx.doi.org/10.1016/j.csl.2016.11.006
0885-2308/ 2016 Published by Elsevier Ltd.

and devising algorithms for automatic pronunciation scoring. Existing works predominantly address phonetic deviances in L2 speech (cf. native speech). For example, Witt and Young (2000) proposed the Goodness of Pronunciation (GOP) based on the likelihood. To achieve phonetic mispronunciation detection and diagnosis (MDD), extended recognition networks (ERNs) based on forced alignment are built to cover the canonical transcriptions as well as some likely error patterns (Ronen et al., 1997; Kawai and Hirose, 1998; Harrison et al., 2009; Wang and Lee, 2012; 2015). Due to the high effectiveness offered by deep learning techniques (Hinton et al., 2006), deep neural networks (DNNs) are also applied to phonetic MDD (Qian et al., 2012; Hu et al., 2013; Lee et al., 2013; Li and Meng, 2014; Li et al., 2016).

With the growing appreciation of suprasegmental training for language learners, more and more efforts are devoted to the research of L2 prosodics, which involves lexical stress (Li et al., 2011a; Li and Meng, 2012; 2013; Li, Meng, 2016), pitch accent (Li et al., 2011a; Zhao et al., 2013b; Li, Meng, 2016), phrasing, rhythm, intonation (Li et al., 2010; Arias et al., 2010a), etc. Lexical stress can be used to disambiguate lexical terms, e.g., "*permit*" versus "*permit*". Pitch accent, which is associated with the prominent syllable within an intonational phrase (IP), usually carries important information (e.g., new, contrastive, uncertain, etc.) and needs attention from the listeners (Wennerstrom, 2001). Phrasing can convey the syntactic structure of an utterance, e.g., disambiguation between continuation versus termination (Meng et al., 2009). Intonation may imply speech acts (e.g., making a statement, asking a question, etc.), or convey the speaker's mood or attitude (Meng et al., 2009).

Previous work (see Section 2) predominantly focused on prosodic evaluation, automatic pitch accent prediction and detection, as well as intonational boundary detection. However, only a few efforts investigated the classification of intonations or edge tones. As far as we know, this work is among the first attempts to develop an automatic intonation classifier for L2 English speech (Li et al., 2010).

There are several challenges impeding the development of L2 intonation classification. First, it is difficult and costly to design, collect and transcribe a suprasegmental corpus for L2 intonation classification. Second, there is no standard annotation convention for intonational pattern transcription. Although the ToBI convention (Beckman and Elam, 1997; Brugos et al., 2006) is widely used in transcribing the intonation and prosodic structure of English speech, it is very detailed and may be very complex for L2 intonation classification. Third, although we know pitch is the feature most related to intonation, how to make use of pitch as the input features for intonation classification is still a problem.

The rest of this paper is organized as follows. Section 2 introduces previous work related to prosodic evaluation, pitch accent prediction and detection, intonation modeling and classification, etc. Section 3 describes the corpus used in this paper. Section 4 develops an automatic intonation classifier using an MD-DNN. Section 5 presents the experimental results of intonation classification. Section 6 summarizes this paper.

2. Related work

In this section, we will introduce the work related to prosodic evaluation, pitch accent prediction and detection, as well as intonation modeling and classification. As we assume that the intonational boundary is already known, we will not cover the topic of intonational boundary detection (Chen et al., 2004; Sridhar et al., 2008; Jeon and Liu, 2009; Ni et al., 2011; Zhao et al., 2013b), which determines whether a specific location (e.g., the end of a syllable or a word) has an intonational boundary or not. In addition, we will introduce the combination of the final pitch accent and following edge tone (FAET), which corresponds closely to a 'nuclear tone' (Pierrehumbert, 1980; Ladd, 2008),

2.1. Automatic prosodic evaluation

Before performing automatic prosodic evaluation, it is important to investigate prosodic features, which usually relate to duration, energy and fundamental frequency (F0). To improve automatic assessment of nativeness of L2 English speech, Teixeira et al. (2000) examined a large set of prosodic features relating to pitch, lexical stress, duration of two longest pauses/words/vowels within the utterance, etc. However, experiments showed that incorporating these prosodic features with segmental features did not improve the performance of nativeness assessment, i.e., the segmental features obtained the best results. To evaluate the naturalness of prosody, Maier et al. (2009) investigated the prosodic features including energy, F0, voiced and unvoiced segments, etc. To evaluate the rhythm and intonation, Suzuki et al. (2008) introduced word importance factors, which is optimized by using a decision tree for word clustering. Jang (2009) and Lai et al. (2013) also exploited many rhythm-related features, e.g., proportion of vowel intervals, proportion of function words, etc. Cucchiarini et al. (2000) and Zechner et al. (2009) investigated various fluency-related features, e.g., speech rate, articulation rate, number and length of pauses, etc.

Yamashita and Nozawa (2005) develop a comparison-based method to evaluate prosodic proficiency of L2 English speech. A native and eight Japanese speakers were asked to utter 60 sentences with various intonation patterns. Other comparison-based approaches for automatic prosodic evaluation include Ito et al. (2006), Ito et al. (2009), Duong et al. (2011), Arias et al. (2010b), etc.

To evaluate the non-native intonation, Tepperman and Narayanan (2008) use Hidden Markov Models (HMMs) to represent intonation units; Ito et al. (2009) combined multiple evaluation scores from multiple decision trees (Suzuki et al., 2008). Other work focus on intonation assessment include Zhao et al. (2010), Duong et al. (2011), and Cheng (2011).

Hönig et al. (2010) applied multiple linear regression on a large prosodic feature vector to assess the quality of L2 learners' utterances with respect to intelligibility, rhythm, melody, etc. The results were further improved by combining these features with those derived from a Gaussian mixture model (GMM), which was used as an universal background model (Hönig et al., 2011). van Santen et al. (2009) proposed approaches for automatic assessment of prosody production, including lexical stress, focus, phrasing, etc. The experimental features were based on spectral, F0 and temporal information.

2.2. Pitch accent prediction and detection

Stressed or accented syllables usually exhibit longer duration, greater loudness and higher pitch than their neighbors (Fry, 1958; Morton and Jassem, 1965; Tamburini, 2003; Li et al., 2011a). Based on these prosodic features, pitch accent can be automatically detected by using HMMs (Imoto et al., 2002; Li et al., 2007), Bayesian classifier assuming multivariate Gaussian distributions (Tamburini, 2003), time-delay recursive neural networks (Ren et al., 2004), support vector machines (SVMs) (Zhao et al., 2013b), multi-layer perceptrons (Zhao et al., 2013a), latent-dynamic conditional neural fields (a kind of probabilistic graphical models) (Tamburini et al., 2014), etc.

Besides prosodic features, lexical and syntactic features also highly correlate with pitch accents. Ross et al. (1992) investigated several factors influencing pitch accent placement based on a subset of the Boston University Radio News Corpus (BURNC) (Ostendorf et al., 1995). They found that 39% of the words were function words and only about 11% of these function words had pitch accents. For the pitch accents on function words, 42% of them were on negatives or quantifiers. Similarly, Rosenberg (2009) showed that about 76% of the content words in the BURNC were accented, whereas only 14% of the function words were accented.

Due to the correspondence with pitch accents, lexical and syntactic features are widely used in automatic pitch accent prediction − a task that assigns pitch accents from given text and is motivated by synthesizing more natural sounding speech. With a set of lexical and syntactic features from unrestricted text, Hirschberg (1993) proposed a method using classification and regression tree (CART) (Breiman et al., 1984; Lewis, 2000) to predict pitch accent location. Experiments showed that part-of-speech (POS) played an important role in the prediction. Ross and Ostendorf (1996) also investigated many kinds of features for pitch accent prediction, including lexical stress, POS, prosodic phrase structure (e.g., phrase break size, number of syllables/words, etc.), new/given status, paragraph structure (e.g., the position of phrase within the sentence, etc.), and labels of other units (e.g., types of pitch accent, boundary tone, etc.).

Furthermore, lexical and syntactic information are also used to complement acoustic features in pitch accent detection (Wightman and Ostendorf, 1994; Conkie et al., 1999; Sun, 2002; Chen et al., 2004; Gregory and Altun, 2004; Levow, 2008; Ananthakrishnan and Narayanan, 2008b; Sridhar et al., 2008; Jeon and Liu, 2009; Qian et al., 2010; Ni et al., 2011).

2.3. Intonation modeling

Fujisaki and Hirose (1982) proposed a model using two critically damped filters to generate the fundamental frequency (F0) contours. The phrase component uses impulses as input and the accent component uses a step function. By specifying the different amplitudes and durations, the model works well for the declarative intonation, yet not so well for gradually rising intonation.

Hirst (1992) developed a model in which the F0 contour is first encoded by a number of target points using a fitting algorithm. It is then classified into different phonological descriptions. Similar to Hirst's model, Taylor (1995) put forward a rise-fall-connection (RFC) model which tries to encode the F0 contour as R(rise), F(fall) and C(connection). After pitch interpolation, smoothing for unvoiced phonemes and perturbations, the F0 contour can be described by rising/falling amplitudes and rising/falling durations, with the assumption that pitch accents and boundaries are explicitly marked.

Taylor (1998); 2006) further proposed a Tilt model, which uses three parameters (i.e., *amplitude, duration* and *tilt*) to describe the intonational shapes of a rise, a fall and a rise followed by a fall. *Amplitude* is defined as the sum of the rise (A_{rise}) and fall (A_{fall}) amplitudes. *Duration* is the sum of the rise (D_{rise}) and fall (D_{fall}) durations. The *tilt* parameter describes the overall intonational shape and is calculated by Eq. (1). In this model, basic events are often associated with vowels.

$$tilt = \frac{A_{\text{rise}} - A_{\text{fall}}}{2(A_{\text{rise}} + A_{\text{fall}})} + \frac{D_{\text{rise}} - D_{\text{fall}}}{2(D_{\text{rise}} + D_{\text{fall}})} \qquad (1)$$

Wright and Taylor (1997) tried to use hidden Markov models (HMMs) to model intonational tunes and automatically identify the tune types of an utterance. Experiments were performed on a subset of the DCIEM Maptask corpus (Bard et al., 1996), which is a goal-directed dialog corpus uttered by Canadian speakers.

2.4. Edge tone prediction and classification

Ross and Ostendorf (1996) used a decision tree to predict the edge tones (L-L%, L-H% or H-L%, see Section 3.2) at the end of each IP. The lexical and syntactic features included punctuation, phrase position, phrase length, etc. Experiments were based on the data of a single speaker in the Boston University Radio News Corpus (BURNC) (Ostendorf et al., 1995), achieving an accuracy of 66.2%. This result was just slightly better than the chance level of 61.1% (setting all edge tones to the majority one, L-L%).

Ananthakrishnan and Narayanan (2008a) proposed a fine-grained boundary tone labeling system. Features were derived from RFC and Tilt models. Experiments were performed on the BURNC, which consists of about three hours of speech from six speakers. The system obtained an accuracy of 67.7%, which was also slightly better than the chance rate of 60.2%.

Rosenberg (2009) investigated the classification of edge tones. The used prosodic features included slope aggregations of pitch (e.g., minimum, maximum, mean, etc.), *tilt* parameters, extrema locations of pitch values. They also used syntactic features including parse tree features (Charniak, 2000) and part-of-speech tags. By examining the parameters over the 200-ms region prior to the intonational phrase boundary, the edge tone classification achieved an accuracy of about 78% based on the BURNC. Experiments also found that the syntactic features were less discriminative than the prosodic features.

2.5. Multi-distribution DNNs (MD-DNNs)

In real applications, input features may have different kinds of distributions, e.g., lexical and syntactic features maybe binary, whereas prosodic features maybe Gaussian. To incorporate these features, Kang et al. (2013) proposed an MD-DNN for speech synthesis, which was also applied to lexical stress detection (Li and Meng, 2013) and phonetic MDD (Li and Meng, 2014). Similar to traditional DNNs, MD-DNNs are also constructed by stacking up multiple Restricted Boltzmann Machines (RBMs) from bottom up. This involves running a layer-by-layer unsupervised pre-training algorithm (Hinton and Salakhutdinov, 2006; Hinton et al., 2006), followed by fine-tuning the pretrained network using the back-propagation algorithm (Rumelhart et al., 1986). Excluding the bottom RBM, all the other ones are traditional Bernoulli RBM, whose hidden and visible units are all binary. The bottom RBM is a type of mixed Gaussian-Bernoulli RBM, whose hidden units are binary while visible units maybe Gaussian or binary.

2.6. Combination of final pitch accent and following edge tone (FAET)

The basic unit of intonation is called the intonational phrase (IP). An IP covers the part of an utterance over which a particular intonation pattern extends, which usually ends at a comma, period, question mark, etc. Basic components of an intonational event include pitch accents and edge tones (Ladd, 2008). Pitch accents associate with syllables to

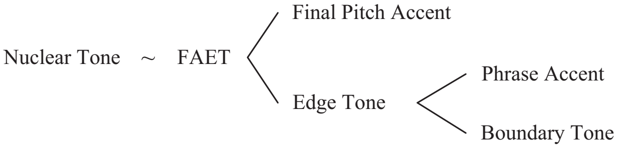


Fig. 1. A nuclear tone corresponds closely to the combination of the final pitch accent and following edge tone (FAET). An edge tone is constituted of a phrase accent and a boundary tone.

signify emphasis; while edge tones occur at the edges of IPs to give cues such as continuation, question or statement. In addition, an edge tone can be further divided into a phrase accent and a boundary tone (Pierrehumbert, 1980; Ladd, 2008), as illustrated in Fig. 1. The combination of the final pitch accent and following edge tone (FAET) corresponds closely to a 'nuclear tone' (Pierrehumbert, 1980; Ladd, 2008), which is important for expressing intonational meaning (Cruttenden, 1997). Hence, we may classify intonation patterns by focusing on the FAET of an IP, i.e., from the final pitch accent to the end of the IP.

3. Experimental corpus

3.1. Supra-CHLOE corpus

We have designed and collected the Supra-CHLOE (**Supra**segmental **Ch**inese **L**earners **O**f **E**nglish) corpus (Li et al., 2011b). It contains speech recordings from 100 Mandarin speakers and 100 Cantonese speakers (both groups are gender-balanced). There are five parts in this corpus: *lexical stress,utterance-level stress, intonation, reduced/ unreduced function words* and *prosodic disambiguation*. The materials used in this corpus are designed by the AESOP (Asian English Speech cOrpus Project) (Visceglia et al., 2009), which is a multinational collaboration aiming to develop a speech corpus that can represent the varieties of English spoken in Asia.

Excluding the *lexical stress* part, all the other parts' syllables are labeled with different types of pitch accents (see Section 3.2). For the *intonation* part, there are 29 sentences covering rising intonation, falling intonation and continuation rises. The speakers were instructed to read in rising tone when "↗" is shown. Similarly, "↘" indicates falling tone. The 29 sentences include four types, as shown in Table 1. Altogether, the 200 speakers recorded 5800 utterances, with 8400 IPs (2200 targeted for rise, 2600 for continuation rise, and 3600 for fall).

3.2. Annotation convention and procedure

A linguistically trained annotator transcribed each IP in two ways: the ToBI convention (Beckman and Elam, 1997; Brugos et al., 2006) for a descriptive labeling of pitch accents and edge tones, and a perceptual judgment for intonation patterns in terms of RULF (Rising/Upper/Lower/Falling). To increase the labeling quality, all the transcriptions were examined by the annotator after about three months. The examination results show that the main variations lied in the labeling of pitch accent.

Table 1
Types of sentences in our corpus and their targeted patterns of the intonational phrase (IP).

Types of sentences	# Sent.	# IPs
Yes−no questions, e.g., Do you need any money ↗?	11	Rise: 11
Wh-questions, e.g., When will John be available↘?	8	Fall: 8
Declarative statements, e.g., In December and January ↗, the sun rises at seven in the morning↘.	8	Cont. Rise: 8 Fall: 8
List-item statements, e.g., He bought strawberries↗, pineapples↗, bananas↗, and apples↘.	2	Cont. Rise: 5 Fall: 2

3.2.1. ToBI convention

We follow the ToBI convention to label pitch accents and edge tones. The major types of pitch accents include H* (peak), L* (low), L+H* (rising peak), L*+H (scoop), H+!H* (falling). In addition, !H*, L+!H*, L*+!H are used where the peak is lower than a preceding high pitch accent; *? is used for uncertainty about whether a pitch accent exists.

For edge tones, there are four types: H-H% (typical yes−no question, rising pitch up to high range), L-H% (list-item intonation, rising pitch, yet not up to high range), L-L% (typical declarative sentence, low edge tone), H-L% (plateau, pitch remain high).

3.2.2. The RULF labels

We also annotate the same set of data using the RULF system with reference to (Selting, 1995). It resembles the British convention (Ladd, 2008) in using **R**ising and **F**alling, and differs by introducing **U**pper and **L**ower. The latter two types are proposed to capture the unclear instances in the L2 English speech uttered by Chinese speakers. This work was previously presented in Li et al. (2010). Examining the pitch contour over the FAET, an IP is first judged whether it is R or F:

- **R**ising: a rising intonation is perceived;
- **F**alling: a falling intonation is perceived.

If no obvious rise or fall can be identified, we will try to determine the pattern as one of the following two types:

- **U**pper: the intonation is perceived as high;
- **L**ower: the intonation is perceived as low.

Finally, if it is still hard to identify an IP as any of the above types, a question mark will be given to indicate uncertainty.

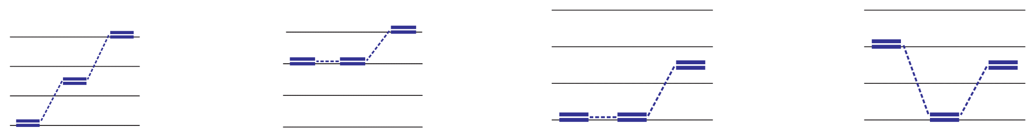
- ? (Question): difficult to classify as R/U/L/F.

3.2.3. Relationship between RULF labels and FAETs

We refer to the **f**inal pitch **a**ccents and their following **e**dge **t**ones as FAETs. Since the RULF labels describe the same part of pitch contour, they correlate closely with FAETs.

- (L*/H*) L-H% and (L*/H*) H-H% may correlate with ‘R’, as all indicate a rising pitch contour (see Fig. 2).
- H* L-L% may correlate with ‘F’; L* L-L% may correlate with ‘F’ or ‘L’ (see Fig. 3).
- L* H-L% may correlate with ‘R’; H* H-L% may correlate with multiple types, namely ‘R’, ‘U’ or ‘F’, depending on the relation between H* and H-, and that between H- and L% (see Fig. 4).

Note that the relationships above are assumed for general patterns and it is possible to observe irregular relationships in real data. In addition, pitch accents like !H*, L+H*, L+!H*, L+H*, L*+!H, and H+!H* are omitted here, as their combination with edge tones may all resemble H* in corresponding to Rising/Upper/Lower/Falling.



(a) L* H-H%: **R** (b) H* H-H%: **R** (c) L* L-H%: **R** (d) H* L-H%: **R**

Fig. 2. Relationship between (L*/H*) L-H%, (L*/H*) H-H% and RULF labels.

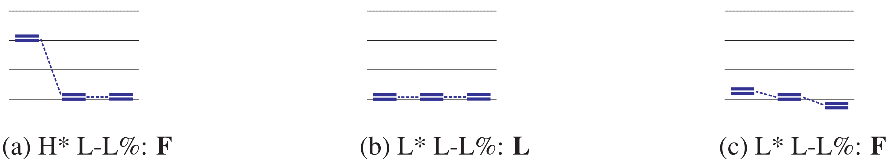

(a) H* L-L%: **F** (b) L* L-L%: **L** (c) L* L-L%: **F**

Fig. 3. Relationship between (L*/H*) L-L% and RULF labels.

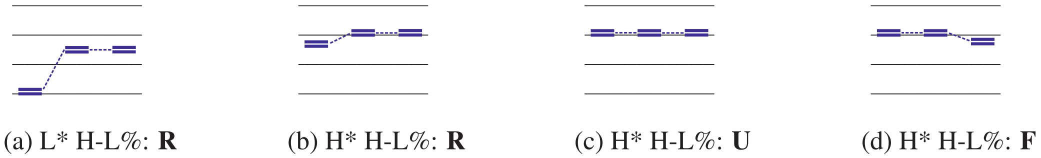

(a) L* H-L%: **R** (b) H* H-L%: **R** (c) H* H-L%: **U** (d) H* H-L%: **F**

Fig. 4. Relationship between (L*/H*) H-L% and RULF labels.

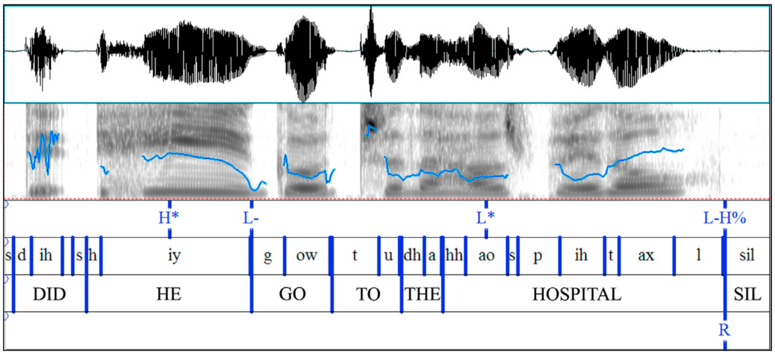


Fig. 5. An example of annotation. From top to bottom, panels show the speech waveform, spectrogram, pitch accents and edge tones, phonemes, words, and RULF labels, respectively.

3.3. Annotation results

Each IP is annotated with pitch accents, edge tones, and RULF labels. An example is shown in Fig. 5. The words and phonemes are indexed by automatic speech recognition.

3.3.1. Pitch accents and edge tones

Table 2 gives the annotation results of pitch accents in rates and counts. Since about 68% of the syllables are transcribed as unaccented, all the syllables are grouped as accented and unaccented in the experiments of pitch accent detection.

Table 3 tabulates the annotation results of edge tones in different types of IPs. It shows that the L2 learners perform best on the IPs targeted for fall, of which 94.2% or (3387 out of 3595) is annotated as L-L%. For the IPs

Table 2
Annotation results of pitch accents in rates and counts. ‘Un’ means unaccented.

H*	!H*	L+H*	L+!H*	H+!H*	L*	L*+H	L*+!H	Un	*?
11.48% (23,594)	4.05% (8311)	9.80% (20,143)	1.28% (2627)	1.88% (3868)	2.70% (5555)	0.73% (1489)	0.06% (122)	68.01% (139,747)	0.01% (19)

Table 3
Distribution of edge tones in different IPs.

Indicated	Annotated				
	L-L%	H-L%	L-H%	H-H%	Total
Fall (↘)	3387	10	193	5	3595
Cont. Rise (↗)	338	186	1785	288	2597
Rise (↗)	93	59	1874	173	2199
Total	3818	255	3852	466	8391

Table 4
Distribution of RULF labels in different IPs.

Indicated	Annotated					
	?	F	L	U	R	Total
Fall (↘)	11	3378	1	4	188	3582
Cont. Rise (↗)	16	340	1	57	2183	2597
Rise (↗)	4	91	0	21	2076	2192
Total	31	3809	2	82	4447	8371

targeted for rise, the L2 learners tend to use L-H% instead of H-H%, whose rates are 85.2% and 7.9% respectively. For the IPs targeted for continuation rise, only 68.7% of these IPs are transcribed as L-H%, and 13.0% as L-L%.

3.3.2. *RULF*

Table 4 shows the annotation results of RULF in different types of IPs. It shows that the L2 learners perform well on the IPs targeted for fall or rise. About 94.3% or (3378 out of 3582) of the IPs targeted for fall is annotated as 'F', and about 94.7% or (2076 out of 2192) of the IPs targeted for rise is transcribed as 'R'. For the IPs of continuation rise, only 84.1% or (2183 out of 2597) is produced with rising intonation, whereas 13.1% or (340 out of 2597) is produced with falling intonation.

3.3.3. *Relationship between RULF labels and FAETs*

The overall distribution of the RULF labels in different kinds of FAETs, from the results of annotation, is given in Table 5. Note that !H* is grouped into H*, regarding their similarity in pitch contour. Similarly, L+!H* and L*+!H are merged into L+H* and L*+H respectively.

Table 5
Distribution of RULF labels in different kinds of FAETs.

FAET	?	F	L	U	R	Total
H* L-L%	1	2353			3	2357
L* L-L%	2	30	2		3	37
L+H* L-L%		1211			6	1217
L*+H L-L%		35			1	36
H+!H* L-L%		153				153
? L-L%		2				2
H* H-L%		1		82	6	89
L* H-L%					83	83
L+H* H-L%	2				76	78
L*+H H-L%					4	4
H+!H* H-L%	1					1
H* L-H%	11	8			1008	1027
L* L-H%	1	4			2199	2204
L+H* L-H%	10	5			506	521
L*+H L-H%					36	36
H+!H* L-H%	3	2			52	57
H* H-H%		1			112	113
L* H-H%		2			286	288
L+H* H-H%					63	63
L*+H H-H%					2	2
Total	31	3807	2	82	4446	8368

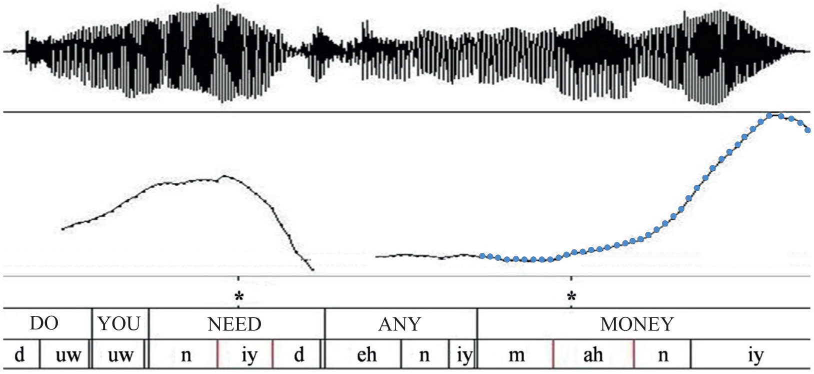


Fig. 6. An example of rising intonation. The pitch accents are located by the pitch accent detector and marked as '*'. The pitch contour over the FAET is highlighted.

Table 5 shows that when the edge tone is L-H% or H-H%, the intonation is primarily annotated as a rising tone 'R', regardless of the final pitch accent. Edge tone L-L% mainly correlates to 'F', regardless of the final pitch accent preceding it. In the case of H-L%, the sequence L* H-L% correlates to 'R'; while H* H-L% mainly correlates to 'U'.

Table 5 also shows that the amount of IPs annotated as 'U' and 'L' is small. Hence, we group 'U' into 'R', for they generally correspond to a rising intonation. Similarly, the cases of 'L' are merged into 'F'. With these processed data, the four-category classification task is simplified to a two-category classification problem.

4. Automatic intonation classification

The automatic intonation classifier focuses on the pitch contour over the FAET, i.e., from the final pitch accent to the end of the IP, which generally corresponds to the location of an orthographic comma, period, question mark, etc. Hence, we should first develop a pitch accent detector which may be used to locate the final pitch accent.

Fig. 6 shows an example with a rising intonation. The sentence is "Do you need any money?". The pitch accents are located by the pitch accent detector and marked as '*', and the pitch contour over the FAET is highlighted. Note that the pitch value used in this work is converted to the semitone scale (Li et al., 2011a) and normalized with the mean pitch value of the IP.

4.1. Pitch accent detection

Similar to the lexical stress detector proposed in our previous work (Li and Meng, 2013), the pitch accent detector is an MD-DNN, whose diagram is shown in Fig. 7. The syllable-based prosodic features include syllable nucleus duration (V_{dur}), maximum loudness (V_{loud}) and a pair of dynamic pitches (f_{m1} and f_{m2}), where f_{m1} and f_{m2} are the first and second extreme pitch values (according to time sequence) in the syllable nucleus respectively. These features are scaled to have zero mean and unit variance over the whole corpus.

In addition, the lexical and syntactic features (PS, SS, NS, NULL and $F_{initial}$) are also used as part of the input features. We use two bits to indicate a syllable carrying primary stress (PS), secondary stress (SS) or no stress (NS) in dictionaries, and an additional bit $F_{initial}$ to indicate an onset of a new word. The NULL means there is no syllable, e.g., for the final syllable in an IP, there are no succeeding syllables. As accented syllables are more prominent than their neighbors, a contextual window of 5 syllables (2 before, 1 current and 2 after) is applied in this work. Thus, there are total 15 binary units in the bottom of the MD-DNN for the lexical and syntactic features, as well as 20 linear units with Gaussian noise for the syllable-based prosodic features.

Above the bottom layer, there are three hidden layers of 128 units. For the top output layer, there are two units generating the posterior probabilities of being accented or unaccented.

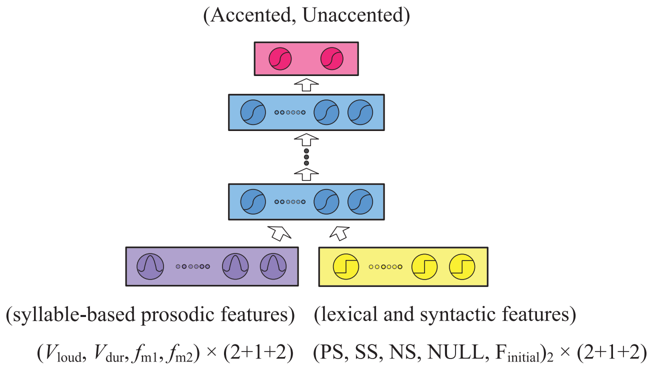

(Accented, Unaccented)
(syllable-based prosodic features)
(lexical and syntactic features)
(V_{loud}, V_{dur}, f_{m1}, f_{m2}) × (2+1+2)
(PS, SS, NS, NULL, $\text{F}_{\text{initial}}$)$_2$ × (2+1+2)

Fig. 7. Diagram of the MD-DNN for pitch accent detection.

4.2. *Implementation of intonation classifier using MD-DNN*

With the pitch accent detector, we can reduce the scope of pitch contour for intonation classification from the whole IP to the FAET. However, we cannot directly use the pitch values of the corresponding FAET, since their length varies with the specific IP. Hence, it is necessary to investigate the distribution of FAETs with different lengths of pitch contour, as shown in Fig. 8. It shows that only 1.7% of the FAETs have a length of pitch contour larger than 128 frames (i.e., 1.28 s). Note that about 3.7% of the FAETs fail in detecting pitch values, which is not shown in the figure.

The diagram of the MD-DNN for intonation classification is shown in Fig. 9. We use N_{FAET} linear units with Gaussian noise to characterize the pitch contour over the FAET. The pitch values for frames of distance greater than N_{FAET} frames from the end of the pitch contour will be truncated. If the pitch contour over the FAET has fewer than N_{FAET} frames, we will fill it with zero value before the onset of the FAET. In addition, two binary nodes are used to indicate that the IP is targeted for fall, continuation rise or rise. Above the bottom layer, there are three hidden layers and each has 64 units. In the top output layer, there are two units generating the posterior probabilities of being a rising or falling intonation.

As 98.3% of the FAETs have a pitch contour whose length is smaller than 128 frames, N_{FAET} should be equal to or smaller than 128. We will analyze the effect of the value of N_{FAET} in our later experiments.

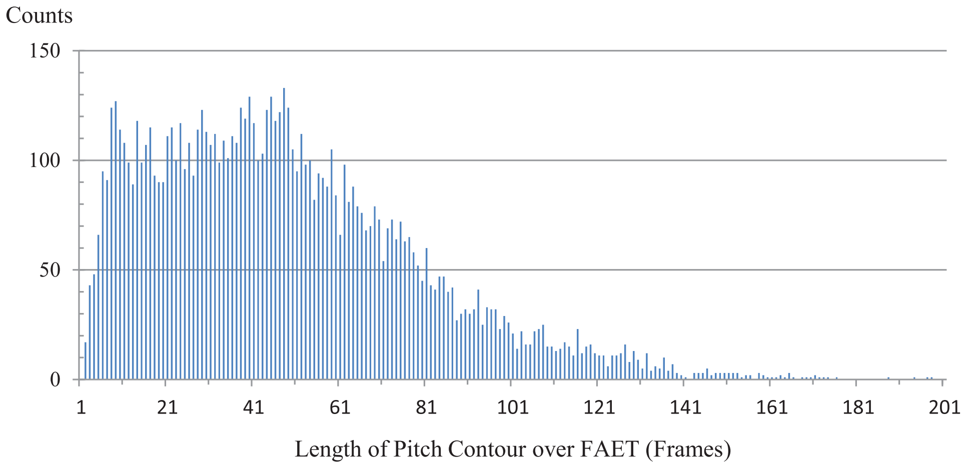


Fig. 8. Counts of FAETs as a function of the length of pitch contour. About 3.7% of the FAETs fail in detecting pitch values, which is not shown here.

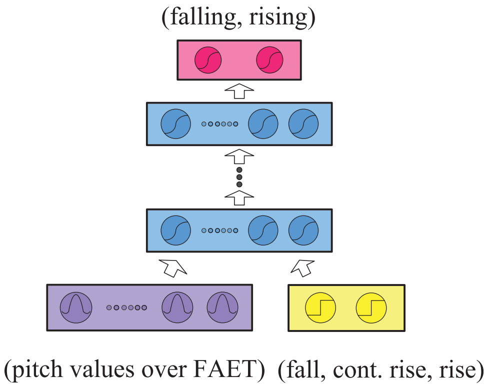


Fig. 9. Diagram of the MD-DNN for intonation classification. In the bottom of the MD-DNN, there are N_{FAET} linear units with Gaussian noise used to characterize the pitch contour over the FAET. The value of N_{FAET} will be determined in later experiments. In addition, two binary units are used to indicate that this IP is targeted for fall, continuation rise or rise.

Table 6
Results of pitch accent detection from 10-fold cross-validation. The syllables in the intonational phrases with only one monosyllabic word are not counted in this experiment.

Detected	Labeled	
	Unaccented	Accented
Unaccented	63.42% (125,952)	5.16% (10,250)
Accented	4.69% (9307)	26.73% (53,091)

5. Experiments

5.1. Pitch accent detection

The evaluation results of the pitch accent detection from 10-fold cross-validation are summarized in Table 6. Among the total 198,600 syllables, 90.15% of them are correctly identified as either accented or unaccented. Among the syllables that are annotated as accented, 83.82% of them are correctly detected.

Our previous work (Li et al., 2011a) adopted Gaussian mixture models (GMMs) for pitch accent detection based on the Supra-CHLOE corpus. Two approaches of detection were investigated: one using the syllable-based prosodic features (see Section 4.1) and the other using the prominence features from the prominence model (PM) (Li and Liu, 2010). The PM estimates the prominence values from the syllable in focus as well as the syllables in neighboring contexts. It is based on the observations that syllables with loudness, duration and pitch greater than their neighboring syllables are likely to be perceived as stressed or accented, even if their values are not large on average. Hence, the differences between the feature values of the current syllable and the ones of the neighboring syllables are also considered. With this PM, the syllable-based prosodic features are converted into a set of prominence features. Note that both approaches were based on supervised learning. For simplicity in notation, we denote the former approach with GMMs and the latter with PM. The results of pitch accent detection using the GMMs, PM and MD-DNN are shown in Fig. 10. We observe that the MD-DNN outperforms the GMMs and PM by about 9.6% and 6.9% respectively.

5.2. Intonation classification

In this section, we first configure the value of N_{FAET} and the MD-DNN structure. Then we examine the contribution of different kinds of features, the influence of pitch accent detection and the performance of using MD-DNNs and SVMs. Finally, we present the detailed experimental results of our intonation classifier from 10-fold cross-validation.

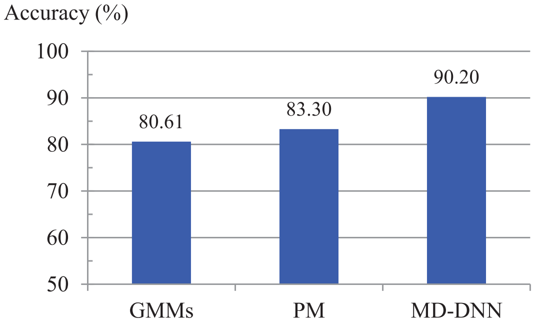


Fig. 10. Performance of pitch accent detection using different classifiers. The results of GMMs and PM are from Li et al. (2011a).

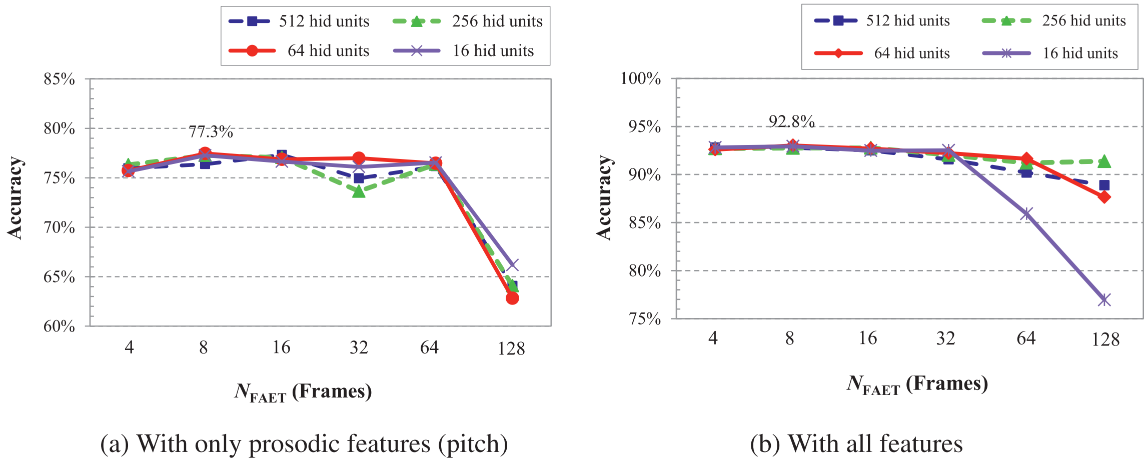
5.2.1. *Configurations of* N*$_{FAET}$ and MD-DNN structure*

To configure the value of N_{FAET} and the structure of MD-DNN, we perform two groups of experiments − one only using prosodic features (i.e., pitch) and one using all kinds of features (i.e., pitch and target intonation patterns). Fig. 11 illustrates the effect of varying the value of N_{FAET} and the size of each hidden layer. The classifiers work quite well when N_{FAET} ranges from 4 to 64. However, the performance would decrease greatly if we use larger values of N_{FAET} (e.g., 128 frames), since there are only 8340 IPs in our corpus and 73.1% of these IPs have a pitch contour with 64 frames or less over the FAET.

All the MD-DNNs in Fig. 11 have three hidden layers. The results show that the MD-DNNs with only 16 hidden units per layer already perform quite well, especially if N_{FAET} is smaller than 64. The intonation classifiers achieve the optimal performance if we use 8 frames as the value of N_{FAET} and 64 units as the size of each hidden layer. These configurations are applied in subsequent experiments. Further increasing the size of hidden units (e.g., 512 units) may cause over-fitting.

5.2.2. *Contribution of different types of features*

Fig. 11 shows that the intonation classification can obtain an accuracy of 77.5% if only prosodic features are leveraged. When we incorporate target intonation patterns, the accuracy can be further improved to 93.0%. Similar improvement can be observed if we use SVMs as the classifier (see Section 5.2.4). This is because 92.5% (i.e., 7716/8340) of the IPs are pronounced following the target indicators (see Table 4).



(a) With only prosodic features (pitch)

(b) With all features

Fig. 11. Performance of intonation classification as a function of the value of N_{FAET} and size of each hidden layer. The MD-DNNs in (a) only use prosodic features (i.e., pitch); whereas the MD-DNNs in (b) using all features (i.e., pitch and target intonation patterns). All the MD-DNNs have three hidden layers. The optimal accuracies for these two figures are labeled.

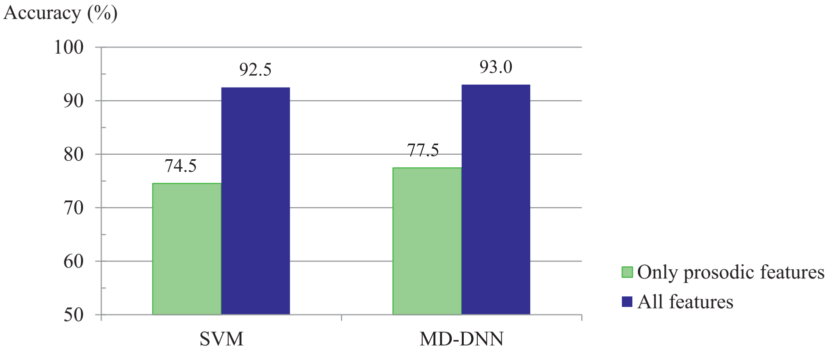


Fig. 12. Performance of intonation classification using SVMs and MD-DNNs.

Table 7
Intonation classification results from 10-fold cross-validation.

Detected	Labeled	
	Falling	Rising
Falling	42.01% (3504)	3.30% (275)
Rising	3.68% (307)	51.01% (4254)

If the target intonation indicators are not presented to non-native speakers, the performance of classifier using both prosodic features and target intonation patterns may decrease greatly, and we may have to rely on the classifier that only uses prosodic features. To overcome this problem in the future, we have to retrain these classifiers by using more non-native speech that the target intonation indicators are not presented.

5.2.3. Contribution of pitch accent detection

If we use a large value for N_{FAET}, it is necessary to use a pitch accent detector to reduce the scope of pitch contour for intonation classification. Without the help of a pitch accent detector (i.e., using the entire final segment of the pitch contour, up to 128 frames, as input features), the accuracy of the MD-DNN with 256 units per hidden layer would be decreased from 91.4% to 83.5%. However, pitch accent detection has little effect on the performance of intonation classification provided we use a small value (e.g., 8 frames) for N_{FAET}.

5.2.4. SVMs versus MD-DNNs

Fig. 12 shows the performance of intonation classification using SVMs and MD-DNNs respectively. If only prosodic features are used, the MD-DNN outperforms the SVM by about 3%. Combining the prosodic features with target intonation patterns, both classifiers obtain similar performance, whose accuracies are 92.5% (SVM) and 93.0% (MD-DNN) respectively.

5.3. Confusion matrix of intonation classification

Fig. 11 illustrates that the intonation classifier obtains the best accuracy if we use all the different kinds of features, 8 frames as the value of N_{FAET} and 64 units as the size of each hidden layer of an MD-DNN with three hidden layers. The detailed evaluation results from 10-fold cross-validation are given in Table 7. It shows that 93.9% of the annotated rising intonation are correctly classified as rising intonation, and 91.9% of the intonation annotated as falling are correctly identified as falling intonation.

6. Conclusions

An intonational phrase (IP) is a basic unit of intonation. In general, an intonation pattern can be determined by the pitch contour from the final pitch accent to the end of the IP. To transcribe speech data for intonation classification, we propose the RULF labels which are used to transcribe an intonation as *rising, upper, lower* or *falling*. These four types of labels can be further merged into two groups − *rising* and *falling*. Based on the annotated data from 100 Mandarin and 100 Cantonese learners, we develop a pitch accent detector and an intonation classifier, both of which use MD-DNNs. The pitch accent detector works similarly as the lexical stress detector and identifies syllables as accented or unaccented with an accuracy of 90.2%. The intonation classifier, which considers only 8 frames (i.e., 80 ms) of pitch value prior to the end of the pitch contour over an IP, determines the intonation of L2 English speech as either rising or falling with an accuracy of 77.5%. If we incorporate target intonation patterns, the accuracy is further improved to 93.0%. If a classified intonation is different from the target one, we consider that a mispronunciation in intonation is detected and the appropriate diagnostic feedback will be provided thereafter.

Acknowledgments

The work is partially supported by the grant from the HKSAR Government Research Grants Council General Research Fund (project number 14207315).

References

Ananthakrishnan, S., Narayanan, S., 2008a. Fine-grained pitch accent and boundary tone labeling with parametric F0 features. In: Proceedings of the 2008 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).

Ananthakrishnan, S., Narayanan, S.S., 2008b. Automatic prosodic event detection using acoustic, lexical, and syntactic evidence. IEEE Trans. Audio Speech Lang. Process. 16 (1), 216–228.

Anderson-Hsieh, J., Johnson, R., Koehler, K., 1992. The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentals, prosody and syllable structure. Lang. Learn. 42, 529–555.

Arias, J.P., Yoma, N.B., Vivanco, H., 2010a. Automatic intonation assessment for computer aided language learning. Speech Commun. 52 (3), 254–267.

Arias, J.P., Yoma, N.B., Vivanco, H., 2010b. Automatic intonation assessment for computer aided language learning. Speech commun. 52 (3), 254–267.

Bard, E.G., Sotillo, C., Anderson, A.H., Taylor, M., 1996. The DCIEM map task corpus: spontaneous dialogue under sleep deprivation and drug treatment. In: Proceedings of the International Conference on Spoken Language Processing (ICSLP).

Beckman, M. E., Elam, G. A., 1997. Guidelines for ToBI labeling, version 3.0.

Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 1984. Classification and Regression Trees. CRC Press.

Brugos, A., Shattuck-Hufnagel, S., Veilleux, N., 2006. Transcribing prosodic structure of spoken utterances with ToBI.

Charniak, E., 2000. A maximum-entropy-inspired parser. In: Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL).

Chen, K., Hasegawa-Johnson, M., Cohen, A., 2004. An automatic prosody labeling system using ANN-based syntactic-prosodic model and GMM-based acoustic-prosodic model. In: Proceedings of the 2008 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).

Cheng, J., 2011. Automatic assessment of prosody in high-stakes english tests. In: Proceedings of Interspeech.

Conkie, A., Riccardi, G., Rose, R.C., 1999. Prosody recognition from speech utterances using acoustic and linguistic based models of prosodic events. In: Proceedings of EUROSPEECH.

Cruttenden, A., 1997. Intonation. Cambridge University Press.

Cucchiarini, C., Strik, H., Boves, L., 2000. Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. J. Acoust. Soc. Am. 107 (2), 989–999.

Duong, M., Mostow, J., Sitaram, S., 2011. Two methods for assessing oral reading prosody. ACM Trans. Speech Lang. Process. 7 (4), 14.

Eskenazi, M., 2009. An overview of spoken language technology for education. Speech Commun. 51 (10), 832–844.

Fry, D.B., 1958. Experiments in the perception of stress. Lang. Speech 1 (2), 126–152.

Fujisaki, H., Hirose, K., 1982. Modelling the dynamic characteristics of voice fundamental frequency with application to analysis and synthesis of intonation. In: Proceedings of International Conference on Interactive Collaborative Learning (ICL).

Gregory, M.L., Altun, Y., 2004. Using conditional random fields to predict pitch accents in conversational speech. In: Proceedings of Association for Computational Linguistics (ACL).

Harrison, A.M., Lo, W.-K., Qian, X.-j., Meng, H., 2009. Implementation of an extended recognition network for mispronunciation detection and diagnosis in computer-assisted pronunciation training. In: Proceedings of Symposium on Languages, Applications and Technologies (SLaTE).

Hinton, G., Osindero, S., Teh, Y., 2006. A fast learning algorithm for deep belief nets. Neural Comput. 18, 1527–1554.

Hinton, G., Salakhutdinov, R.R., 2006. Reducing the dimensionality of data with neural networks. Science 313, 504–507.

Hirschberg, J., 1993. Pitch accent in context predicting intonational prominence from text. Artif. Intell. 63 (1), 305–340.

Hirst, D., 1992. Talking machines: theories, models, and designs. Elsevier, North-Holland, Amsterdam, pp. 77–82.

Hönig, F., Batliner, A., Nöth, E., 2011. How many labellers revisited − naives, experts, and real experts. In: Proceedings of InterspeecH.

Hönig, F., Batliner, A., Weilhammer, K., Nöth, E., 2010. Automatic assessment of non-native prosody for English as l2. In: Proceedings of Speech Prosody.

Hu, W., Qian, Y., Soong, F., 2013. A new DNN-based high quality pronunciation evaluation for computer-aided language learning (CALL). In: Proceedings of Interspeech.

Imoto, K., Tsubota, Y., Raux, A., Kawahara, T., Dantsuji, M., 2002. Modeling and automatic detection of English sentence stress for computer-assisted English prosody learning system. In: Proceedings of the International Conference on Spoken Language Processing (ICSLP).

Ito, A., Konno, T., Ito, M., Makino, S., 2009. Evaluation of english intonation based on combination of multiple evaluation scores. In: Proceedings of Interspeech.

Ito, A., Nagasawa, T., Ogasawara, H., Suzuki, M., Makino, S., 2006. Automatic detection of English mispronunciation using speaker adaptation and automatic assessment of English intonation and rhythm. Educ. Technol. Res. 29 (1), 13–23.

Jang, T.-Y., 2009. Automatic assessment of non-native prosody using rhythm metrics: focusing on Korean speakers' English pronunciation. In: Proceedings of International Conference on Economics of Arts and Literature (ICEAL).

Jeon, J.H., Liu, Y., 2009. Automatic prosodic events detection using syllable-based acoustic and syntactic features. In: Proceedings of the 2008 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).

Kang, S., Qian, X., Meng, H., 2013. Multi-distribution deep belief network for speech synthesis. In: Proceedings of the 2008 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).

Kawai, G., Hirose, K., 1998. A method for measuring the intelligibility and nonnativeness of phone quality in foreign language pronunciation training. In: Proceedings of the International Conference on Spoken Language Processing (ICSLP).

Ladd, D.R., 2008. Intonational Phonology. Cambridge University Press.

Lai, C., Evanini, K., Zechner, K., 2013. Applying rhythm metrics to non-native spontaneous speech.. In: Proceedings of Symposium on Languages, Applications and Technologies (SLaTE).

Lee, A., Zhang, Y., Glass, J., 2013. Mispronunciation detection via dynamic time warping on deep belief network-based posteriorgrams. In: Proceedings of the 2008 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).

Levow, G.-A., 2008. Automatic prosodic labeling with conditional random fields and rich acoustic features. In: Proceedings of International Joint Conference on Natural Language Processing (IJCNLP).

Lewis, R.J., 2000. An introduction to classification and regression tree (CART) analysis. In: Proceedings of Annual Meeting of the Society for Academic Emergency Medicine in San Francisco, California, pp. 1–14.

Li, C., Liu, J., Xia, S., 2007. English sentence stress detection system based on HMM framework. Appl. Math. Comput. 185 (2), 759–768.

Li, K., Liu, J., 2010. English sentence accent detection based on auditory features. J. Tsinghua Univ. Sci. Technol. 50 (4), 613–617.

Li, K., Meng, H., 2012. Perceptually-motivated assessment of automatically detected lexical stress in L2 learners' speech. In: Proceedings of International Symposium on Chinese Spoken Language Processing (ISCSLP).

Li, K., Meng, H., 2013. Lexical stress detection for L2 English speech using deep belief networks. In: Proceedings of Interspeech.

Li, K., Meng, H., 2014. Mispronunciation detection and diagnosis in L2 English speech using multi-distribution deep neural networks. In: Proceedings of International Symposium on Chinese Spoken Language Processing (ISCSLP).

Li, K., Meng, H., 2016. Automatic lexical stress and pitch accent detection for L2 English speech using multi-distribution deep neural networks. Speech Commun., (forthcoming).

Li, K., Qian, X., Meng, H., 2016. Mispronunciationdetection and diagnosis in L2 English speech using multi-distribution deep neural networks. IEEE/ACM Trans. Audio Speech Lang. Process. doi: 10.1109/TASLP.2016.2621675.

Li, K., Zhang, S., Li, M., Lo, W.-K., Meng, H., 2010. Detection of intonation in L2 English speech of native Mandarin learners. In: Proceedings of International Symposium on Chinese Spoken Language Processing (ISCSLP).

Li, K., Zhang, S., Li, M., Lo, W.-K., Meng, H., 2011a. Prominence model for prosodic features in automatic lexical stress and pitch accent detection. In: Proceedings of Interspeech.

Li, M., Zhang, S., Li, K., Harrison, A., Lo, W.-K., Meng, H., 2011b. Design and collection of an L2 English corpus with a suprasegmental focus for Chinese learners of English. In: Proceedings of the International Congress of Phonetic Sciences (ICPhS).

Maier, A.K., Hönig, F., Zeiler, V., Batliner, A., Körner, E., Yamanaka, N., Ackermann, P., Nöth, E., 2009. A language-independent feature set for the automatic evaluation of prosody.. In: Proceedings of Interspeech.

Meng, H., Tseng, C.-Y., Kondo, M., Harrison, A., Viscelgia, T., 2009. Studying L2 suprasegmental features in asian Englishes: a position paper. In: Proceedings of Interspeech.

Morton, J., Jassem, W., 1965. Acoustic correlates of stress. Lang. Speech 8 (3), 159–181.

Ni, C.-J., Liu, W., Xu, B., 2011. Automatic prosodic events detection by using syllable-based acoustic, lexical and syntactic features. In: Proceedings of Interspeech.

Ostendorf, M., Price, P.J., Shattuck-Hufnagel, S., 1995. The BostonUniversity radio news corpus. Linguistic Data Consortium 1–19.

Pierrehumbert, J.B., 1980. The Phonology and Phonetics of English inztonation. Massachusetts Institute of Technology Ph.D. thesis.

Qian, X.-j., Meng, H., Soong, F., 2012. The use of DBN-HMMs for mispronunciation detection and diagnosis in L2 English to support computer-aided pronunciation training. In: Proceedings of Interspeech.

Qian, Y., Wu, Z., Ma, X., Soong, F., 2010. Automatic prosody prediction and detection with conditional random field (CRF) models. In: Proceedings of International Symposium on Chinese Spoken Language Processing (ISCSLP).

Ren, Y., Kim, S.-S., Hasegawa-Johnson, M., Cole, J., 2004. Speaker-independent automatic detection of pitch accent. In: Proceedings of Speech Prosody.

Ronen, O., Neumeyer, L., Franco, H., 1997. Automatic detection of mispronunciation for language instruction. In: Proceedings of EUROSPEECH.

Rosenberg, A., 2009. Automatic Detection and Classification of Prosodic Events. Columbia University Ph.D. thesis.

Ross, K., Ostendorf, M., 1996. Prediction of abstract prosodic labels for speech synthesis. Comput. Speech Lang. 10 (3), 155–185.

Ross, K., Ostendorf, M., Shattuck-Hufnagel, S., 1992. Factors affecting pitch accent placement. In: Proceedings of the International Conference on Spoken Language Processing (ICSLP).

Rumelhart, D.E., Hinton, G., Williams, R.J., 1986. Learning representations by back-propagating errors. Natrue 323, 533–536.

van Santen, J.P., Prud'hommeaux, E.T., Black, L.M., 2009. Automated assessment of prosody production. Speech Commun. 51 (11), 1082–1097.

Selting, M., 1995. Prosodie im Gespräch: Aspekte einer interaktionalen Phonologie der Konversation, vol.329. Walter de Gruyter.

Sridhar, V.R., Bangalore, S., Narayanan, S.S., 2008. Exploiting acoustic and syntactic features for automatic prosody labeling in a maximum entropy framework. IEEE Trans. Audio Speech Lang. Process. 16 (4), 797–811.

Sun, X.-J., 2002. Pitch accent prediction using ensemble machine learning. In: Proceedings of the International Conference on Spoken Language Processing (ICSLP).

Suzuki, M., Konno, T., Ito, A., Makino, S., 2008. Automatic evaluation system of english prosody based on word importance factor. J. Syst. Cybern. Inform. 6 (4), 83–90.

Tamburini, F., 2003. Prosodic prominence detection in speech. In: Proceedings of Signal Processing and its Applications.

Tamburini, F., Bertini, C., Bertinetto, P.M., 2014. Prosodic prominence detection in Italian continuous speech using probabilistic graphical models. In: Proceedings of Speech Prosody.

Taylor, P., 1995. The rise/fall/connection model of intonation. Speech Commun. 15, 169–186.

Taylor, P., 1998. The Tilt intonation model. In: Proceedings of ISCLP.

Taylor, P., 2006. Analysis and synthesis of intonation using the tilt model. J. Acoust. Soc. Am. 107, 1697–1714.

Teixeira, C., Franco, H., Shriberg, E., Precoda, K., Sönmez, M.K., 2000. Prosodic features for automatic text-independent evaluation of degree of nativeness for language learners. In: Proceedings of Interspeech.

Tepperman, J., Narayanan, S.S., 2008. Better nonnative intonation scores through prosodic theory. In: Proceedings of Interspeech.

Visceglia, T., Tseng, C.-y., Kondo, M., Meng, H., Sagisaka, Y., 2009. Phonetic aspects of content design in aesop (asian english speech corpus project). In: Proceedings of the 2009 Oriental COCOSDA International Conference on Speech Database and Assessments.

Wang, Y., Lee, L., 2015. Supervised detection and unsupervised discovery of pronunciation error patterns for computer-assisted language learning. IEEE Trans. Audio Speech Lang. Process. 23, 564–579.

Wang, Y.-B., Lee, L.-S., 2012. Improved approaches of modeling and detecting error patterns with empirical analysis for computer-aided pronunciation training. In: Proceedings of the 2008 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).

Wennerstrom, A., 2001. The Music of Everyday Speech: Prosody and Discourse Analysis. Oxford University Press.

Wightman, C.W., Ostendorf, M., 1994. Automatic labeling of prosodic patterns. IEEE Trans. Speech Audio Process. 2 (4), 469–481.

Witt, S.M., Young, S.J., 2000. Phone-level pronunciation scoring and assessment for interactive language learning. Speech Commun. 30 (2), 95–108.

Wright, H., Taylor, P. A., 1997. Modelling intonational structure using hidden Markov models.

Yamashita, Y., Nozawa, K., 2005. Automatic scoring for prosodic proficiency of English sentences spoken by japanese based on utterance comparison. IEICE Trans. Inf. Syst. 88 (3), 496–501.

Zechner, K., Higgins, D., Xi, X., Williamson, D.M., 2009. Automatic scoring of non-native spontaneous speech in tests of spoken English. Speech Commun. 51 (10), 883–895.

Zhao, J., Xu, J., Zhang, W.-q., Yuan, H., Liu, J., Xia, S., 2013a. Exploiting articulatory features for pitch accent detection. J. Zhejiang Univ. Sci. C 14 (11), 835–844.

Zhao, J., Zhang, W.-Q., Yuan, H., Johnson, M.T., Liu, J., Xia, S., 2013b. Exploiting contextual information for prosodic event detection using auto-context. EURASIP J. Audio Speech Music Process. 2013 (1), 1–14.

Zhao, S., Luke, K.K., Koh, S., Zhang, Y., 2010. Computer aided evaluation of intonation for language learning based on prosodic unit segmentation. In: Proceedings of Annual Summit and Conference of Asia-Pacific Signal and Information Processing Association (APSIPA).