Detection of Intonation in L2 English Speech of Native Mandarin Learners

Kun LI, Shuang ZHANG, Mingxing LI, Wai-Kit LO and Helen MENG
Human-Computer Communications Laboratory
The Chinese University of Hong Kong, Hong Kong
{kli, zhangs, mxli, wklo, hmmeng}@se.cuhk.edu.hk

Abstract — **We aim to detect salient mispronunciations in intonation of English speech uttered by Mandarin speakers. The goal of our project is to detect intonation errors and provide corrective feedback to English second language (ESL) learners. An intonational event includes the pitch accent and edge tone, and the intonation is closely related to the nuclear tone of an intonational phrase (IP). Hence, we first develop a pitch accent detector to delineate the scope of analysis in an utterance. Then we develop a nuclear tone detector to classify the intonation of the IP as either rising or falling. The pitch accent detector is a Gaussian mixture model using the features based on energy, pitch contour and the duration of the vowels. The intonation detector is a Gaussian discriminator using three features derived from the pitch contour. Annotated L2 English speech from 40 Mandarin speakers is used in a 10-fold cross-validation setting. The pitch accent detector achieves an accuracy of 72.86%, while its EER is 33.00%. The average classification performance of the intonation detector is 91.17% in accuracy and the EER is 8.60%.**

***Keywords-** language learning, English intonation, ESL learners, L2 suprasegmental features*

I. INTRODUCTION

Prosody of speech plays an important role in the determination of proficiency, intelligibility and is also used in the resolution of semantic ambiguity. Proper placement of prosodic features (e.g., pause, pitch, energy) can help deliver the intended message appropriately. In addition, prosody is also useful in other aspects such as signifying word emphasis, identifying speech acts etc. For an English second language (ESL) learner, mastering the prosody can improve the intelligibility and perceived proficiency of their spoken English. In comparison with segmental phonology (i.e. phonetics), perceptual studies has shown that suprasegmentals (i.e. prosody) may have a stronger effect [1].

Language transfer occurs both in phonetics and prosody. We often observe characteristics of the primary language (L1) in the second language (L2) under acquisition. For example, Mandarin is a syllable-timed language and English is stress-timed. Chinese learners may read English with a syllable-timed rhythm.

A previous investigation [2] has shown that ESL learners demonstrate good perceptual ability in suprasegmental features. The major problem lies in the lack of prior knowledge in the proper use of suprasegmental features. Therefore, a key priority in training Chinese ESL learners is to enrich their knowledge. In the context of computer-assisted pronunciation training (CAPT), the system should be able to pinpoint the salient suprasegmental problems of the learners and provide corrective feedback for their improvement. As a first step, this work focuses on detection of intonation patterns in L2 English speech by Chinese learners.

II. BACKGROUND

Intonation models, such as the Fujisaki model [3], Hirst model [4], rise/fall/connection (RFC) model [5] and Tilt model [6], aim to provide linguistically meaningful interpretations to the acoustics of an utterance. The Fujisaki model uses two critically damped filters to generate the F0 contours: the phrase component uses impulses as input and the accent component uses a step function. By specifying different amplitudes and durations, the model works well for the declarative intonation, yet not so well for gradually rising intonation. In the Hirst model, 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 the Hirst model, the RFC method 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. In the Tilt model, amplitude, duration and Tilt itself are used for describing the intonation shapes of rise, fall and a rise followed by a fall. In this model, a basic event is often associated with vowels.

Basic components of an intonational event include pitch accents and edge tones [7]. Pitch accents associate with syllables to signify emphasis while edge tones occur at the edges of the intonational phrase (IP) to give cues such as continuation, question or statement. An IP covers the part of an utterance over which a particular intonation pattern extends, which usually ends at orthographic commas, periods, or question marks, etc. in the corresponding written sentences. A nuclear tone is defined as the combination of the nuclear pitch accent, i.e. the final pitch accent in an IP, and the edge tone [7]. As illustrated in Fig. 1, the edge tone can further be divided into a phrase accent and boundary tone [7] [8].

In this work, we aim to detect intonation patterns by focusing on the nuclear tone, which is important for expressing intonational meaning [9]. We first find the location of the nuclear pitch accent using a pitch accent detector. We then

ISBN 978-1-4244-6245-2/10/$26.00 ©2010 IEEE

characterize the F0 contour between the detected nuclear pitch accent and the end of the edge tone and classify the L2 learners' intonation as either *rising* or *falling*.

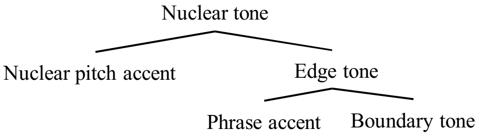


Figure 1. Structure of the nuclear tone.

III. Experimental Data

The data used in this work is a set of 20 male and 20 female speakers from an L2 English speech corpus read by native Mandarin speakers. The speakers were asked to read 29 prompted sentences and instructed to read with a rising or falling intonation, according to an indicator next to each sentence (↗ as rising, or ↘ as falling).

The 29 sentences include four types, as shown in TABLE I. There are 11 IPs targeted for rise, 18 for fall and 13 for continuation rise. Altogether, the 40 speakers recorded 1160 utterances, with 1680 IPs (440 targeted for rise, 520 for continuation rise, and 720 for fall).

TABLE I. Type of the prompting sentences and the targeted patterns of the intonation phrase (IP).

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

IV. The Annotation Procedure

Each IP is annotated in two ways: a *descriptive* labeling of pitch accents and edge tones, and a *perceptual* judgment in terms of RULF (see IV.B below).

A. Pitch accents and edge tones

Following the ToBI convention in [10] and [11], we annotated the following types of pitch accents and edge tones.

1) Pitch accents: H* (peak), L* (low), L+H* (rising peak), L*+H (scoop), H+!H* (falling) (see Fig. 2). 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.

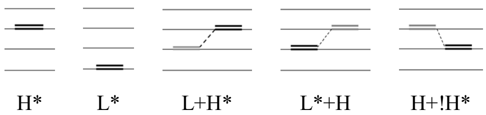


Figure 2. Types of pitch accents (darker lines indicate the prominent tonal target, lighter lines indicate the leading/trailing tones)

2) Edge tones: 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), see Fig. 3.

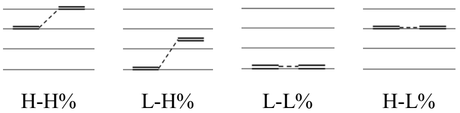


Figure 3. Four types of edge tones

B. The RULF labels

With reference to [12], we annotate the same set of data data by RULF system, a perceptual judgment of the ESL speakers' intonation. RULF resembles the British convention [7] in using R(ising) and F(alling), and differs by introducing U(pper) and L(ower), two types proposed to capture the unclear instances in the L2 English speech by the Chinese speakers.

Examining the pitch contour from the nuclear pitch accent to the end, 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, the annotator will try to identify 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.

C. Correlation between nuclear tones and RULF

As the nuclear tone (nuclear pitch accent plus edge tone) and RULF describe the same part of pitch contour, we assume certain correlation between them, following [7]:

- (L*/H*) L-H% and (L*/H*) H-H% may correlate with **R**(ising), as all indicate rising pitch contour (see Fig. 4):

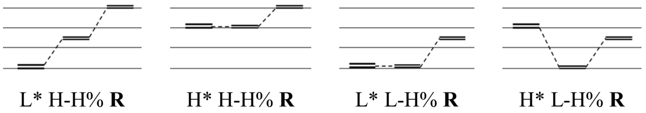


Figure 4. Correlation between (L*/H*) L-H%, (L*/H*) H-H% and RULF

- H*L-L% may correlate to **F**(alling); L*L-L% may correlate with F(alling) or L(ower) (see Fig. 5):

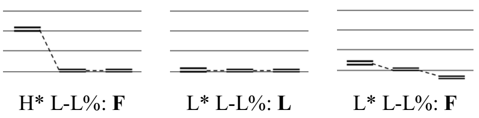
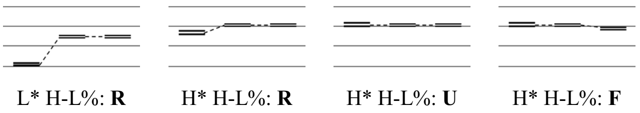


Figure 5. Correlation between (L*/H*) L-L% and RULF

- L*H-L% may correlate to R(ising); H*H-L may correlate to more than one type, namely R(ising), U(pper) or F(alling), depending on the relation between H* and H-, and that between H- and L% (see Fig. 6):



L* H-L%: **R** H* H-L%: **R** H* H-L%: **U** H* H-L%: **F**

Figure 6. Correlation between (L*/H*) H-L% and RULF

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

V. Annotation Results

Each IP is annotated with pitch accents, edge tones, and RULF. An example is shown in Fig. 7. The words and phonemes are indexed by the automatic speech recognition.

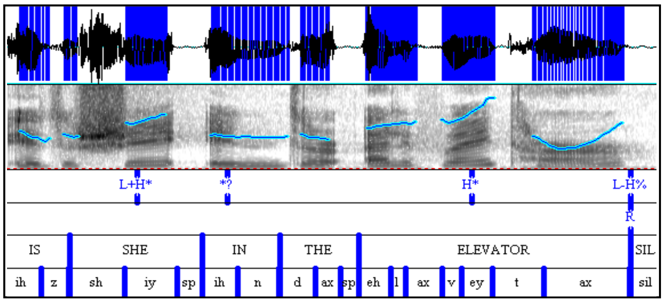


Figure 7. An example annotation

A. Pitch accents

TABLE II gives the annotation results of syllables in the 1160 recorded utterances (1680 IPs), either as an unaccented syllable or a pitch-accented syllable.

TABLE II. Annotation results of pitch accents in counts and rates(%), 'Un' means unaccented

Un	H*	!H*	L*	L+H*	L*+H	L+!H*	L*+!H	H+!H*	* ?
10779	1885	604	947	366	47	45	1	247	444
70.15	12.27	3.93	6.16	2.38	0.31	0.29	0.01	1.61	2.89

B. Edge tones

TABLE III tabulates the annotation results of edge tones in different types of IPs. It shows that the ESL learners perform best on the IPs targeted for fall, of which 99% (709/720) are annotated as L-L%. For IPs targeted rise and continuation rise, H-H% or L-H% are both used, i.e. 207:183 for rise and 160:244 for continuation rise. This indicates that the ESL learners tend to use H-H% versus L-H% interchangeably.

TABLE III. Distribution of edge tones in different IPs

Annotation Indicated Intonation	H-H%	L-H%	H-L%	L-L%	Total
Rise (↗)	207	183	28	22	440
Cont. Rise (↗)	160	244	49	67	520
Fall (↘)	3	7	1	709	720
Total	370	434	78	798	1680

C. RULF

TABLE IV shows the annotation results of RULF in different types of IPs. It shows that the ESL learners perform best on the IPs targeted for fall, 98% (705/720) of which are annotated as 'F'. For IPs targeted for rise, the performance comes close at 90% (394/440). For IPs of continuation rise, only 80% (418/520) are produced with rising intonation, whereas 12% (64/520) are produced with falling intonation.

TABLE IV. Distribution of RULF labels in different IPs

Annotation Indicated intonation	R	U	L	F	?	Total
Rise (↗)	394	19	5	17	5	440
Cont. Rise (↗)	418	28	1	64	9	520
Fall (↘)	11	0	2	705	2	720
Total	823	47	8	786	16	1680

D. Correlation between nuclear tones and RULF

The overall correlation between RULF and nuclear tones, from the result of annotation, is given in TABLE V. Note that !H* is grouped into H*, regarding their similarity in pitch contour. Similarly, L+!H* is grouped into L+H*; and L*+!H is grouped into L*+H.

TABLE V. Correlation between nuclear tones and RULF

Nuclear tones	R	U	L	F	?	Sum
H* H-H%	39	3				42
L* H-H%	314					314
L+H* H-H%	5					5
*? H-H%	9					9
H* H-L%	3	40				43
L* H-L%	21	2				23
L+H* H-L%	10					10
*? H-L%		2				2
H* L-H%	59			1	9	69
L* L-H%	305				1	306
L+H* L-H%	25				1	26
H+!H*L-H%	6				1	7
*? L-H%	26					26
H* L-L%	1		1	661	2	665
L* L-L%			7	38	1	46
L+H* L-L%				67	1	68
L*+H L-L%				1		1
H+!H*L-L%				5		5
*? L-L%				13		13
Total	823	47	8	786	16	1680

Note: the nuclear tones of L*+H H-H%, H+!H* H-H%, L*+H H-L%, H+!H* H-L%, L*+H L-H% are omitted, as these combinations did not appear in the annotation.

TABLE V shows that the correlation between the nuclear tones and the RULF labels generally confirms the correspondence assumed in IV.C: when the edge tone is L-H% or H-H%, the intonation is primarily annotated as a rising tone 'R', regardless of the nuclear pitch accent; edge tone L-L% mainly correlates to 'F', regardless of the nuclear pitch accent preceding it; in the case of H-L%, the nuclear tone L* H-L% correlates to R, while H* H-L% partly correlates to 'U' and in a few cases correlates to 'R'.

VI. Pitch Accent Detector

As mentioned before, the first step of intonation detection is to identify the range of the utterance that maps to the nuclear tone. Therefore, we have developed a pitch accent. The location of the nuclear (final) pitch accent marks the beginning of the nuclear tone.

A. *Acoustic features*

Prosodic prominence involves two different phonetic features: pitch accents and stress. It is also shown that the accented syllables may exhibit a longer duration, greater energy and higher pitch [13]. Therefore, the features used in the pitch accent detector are based on duration, energy and pitch.

1) Vowel normalized duration: It was shown in [13] that the use of the entire syllable duration and the nucleus duration in prominence detection give almost the same performance. In this work, we selected the vowel duration which is normalized by subtracting the mean duration of all vowels in the utterance.

2) Maximum normalized energy: For the energy feature, we first obtain the root-mean-squares (RMS) of the amplitude in log-scale for each frame (10 ms), as in (1):

$$E(n) = 10\log_{10}\left(\frac{1}{N}\sum_{k=1}^{N} A_k^2\right) \qquad (1)$$

where E(n) is the energy of the nth-frame, N is the number of the samples per frame, A_k is the amplitude of the k^{th} sample in the n^{th} frame of the speech.

The maximum frame energy within a vowel [14], which is normalized by subtracting the mean energy over all vowels in the utterance, is used as one of the detector's feature.

3) Maximum normalized pitch: We use the maximum pitch value in a vowel, which provides the most discriminating information for prominence detection among the maximum, minimum, mean, range, etc. of the pitch value in a syllable [14], as the detector's feature. The pitch detection is based on the Snack Sound Toolkit [15]. A post-processing procedure is also applied to smooth out variations, such as octave-jumps.

The detected pitch values are transformed to a log-scale to better match with human perception. We make use of the semitone scale [16], as in (2):

$$f = 12\log_2(f_0 / f_{bottom}) \qquad (2)$$

where f_0 is the pitch in Hertz, f_{bottom} is a normalization factor in Hertz scale obtained using a "90% criterion" (90% of all pitch values in the utterance fall above f_{bottom}), and f is the pitch value in semitone scale. As an example, the f_{bottom} in an utterance is shown in Fig. 10.

B. *Classifier*

In order to discriminate between accented and unaccented syllables, a two-category Gaussian mixture model (GMM) using all of the above 3 features is built. The model for unaccented syllables has only one mixture component. The model for accented syllables has two mixture components – during training, one of the mixtures is trained with high pitch accent data, including the syllables labeled as H*, !H*, L+H*, L*+H, L+!H*, L*+!H, or H+!H*, and the other mixture is trained with low pitch accent data, i.e. the syllables labeled as L*.

VII. Automatic Intonation Detector

As mentioned before, the automatic intonation detector focused on the pitch contour over the nuclear tone, i.e., from the nuclear pitch accent (located by the pitch accent detector) to the end of the IP (corresponding to the locations of orthographic commas, periods, question marks, etc).

A. *Patterns of rising and falling intonation*

All intonation patterns will be recognized as either rising or falling, which is simplified and illustrated in Fig. 8. The points of f_1 and f_2 are the maximum and minimum values in the pitch contour over the nuclear tone, and f_3 is the phrase-ending pitch value. At present, our methodology does not yet consider the duration between these features.

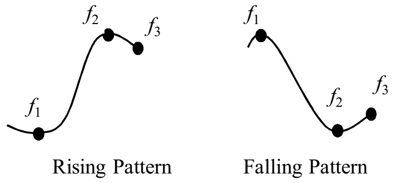


Figure 8. The Rising and Flling Patterns of intonation

Fig. 9 shows the special cases of the risng/falling patterns of intonation. In Fig. 9(a) and Fig. 9(b) the values of f_2 and f_3 are the same.

When the pitch accent detector fails to correctly detect the nuclear pitch accent, or when the nuclear tone is too complicated, we may have cases shown in Fig. 9(c, d), i.e. the difference between f_3 and f_2 is too large, e.g. larger than 2 semitones. In these cases, f_1' and f_2', are used as the detector's features in place of f_1 and f_2.

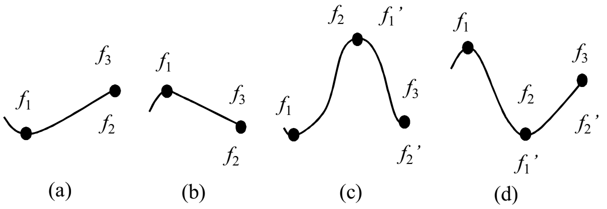


Figure 9. Special cases of intonation patterns. In (a) and (b), f_1, f_2, and f_3 are the 3 features for the intonation detector. In (c) and (d), f_1', f_2', and f_3 are the 3 features for the intonation detector.

As there are strong correspondences between nuclear tones and RULF, and RULF can be simply taken as rising or falling, there are also strong correspondence between nuclear tones and the rising/falling patterns. For example, L*H-H%, H*H-H%, L*L-H% generally correspond to Fig. 9(a), which is a rising pattern; H*L-H% may corresponds to Fig. 9(d), which is also a rising pattern; H*L-L% corresponds to the falling pattern in Fig. 8; L*H-L% corresponds to the rising pattern in Fig. 8; L*L-L may correspond to Fig. 9(b), which is a falling pattern; and H*H-L% may correspond to the rising pattern.

Fig. 10 is an example from the corpus which has a rising tone (The sentence is "Do you need any money?"). The pitch accents are located by the pitch accent detector and marked as '*'. The f_{bottom} for normalization is shown as the dash line. f_1, f_2 and f_3 are the features used in the intonation detector.

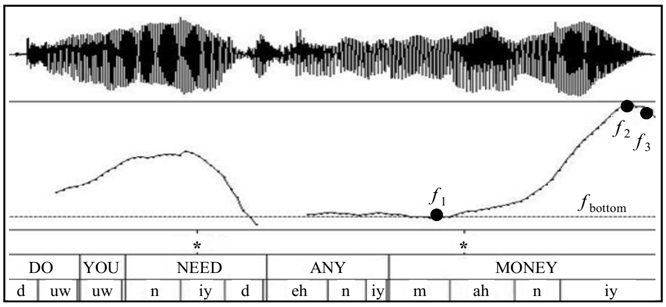


Figure 10. An example of rising intonation. The f_{bottom} for this utterance is indicated by the dash line.

B. *Classifier*

As shown in TABLE IV, the amount of IPs annotated as 'U' and 'L' is small. When we develop the intonation detector, we grouped 'U' into 'R', for they generally correspond to a rising intonation. For the cases of 'L', they generally correspond to a falling intonation and hence are merged into 'F'. By using these processed data, the intonation detection task is simplified to a two-category classification problem for either rising or falling intonation.

Using the features of f_1, f_2, and f_3, a multivariate two-category Gaussian discriminator, i.e. GMM with only one component mixture, is built for the intonation detector which can classify the IPs as rising or falling intonation.

VIII. Experiments and Analysis

We performed the evaluation of the intonation detector using the cross-validation method. The 40 speakers are randomly partitioned into 10 sub-groups (each subgroup has data from 2 females and 2 males). The evaluation is repeated 10 times, every time using one of the 10 sub-groups as the test set and the remaining 9 sub-groups as the training set.

A. *Performance of the pitch accent detector*

The evaluation results of the pitch accent detector are summarized in TABLE VI. The pitch accent detector correctly identified 72.99% of the syllables as either accented or unaccented, and the average Equal Error Rate (EER)[1] for training data across the 10-fold cross-validation is 32.04%.

TABLE VI. Pitch Accent Detection Results

Anno. / Rec.	H*	L*	Unaccented	?
Accented	604	102	594	44
Unaccented	2543	893	10185	400

Note: the cases of "*?" are omitted in calculating the accuracy.

TABLE VI also shows that only 17.04% of the accented syllables, i.e., H* or L*, can be detected, as there are about 70% of the syllables are unaccented (shown in TABLE II), and thus the priors are skewed. For many IPs, no accented syllable can be detected – there are about 1680 IPs and only 706 (604+102) detected pitch accents.

To overcome the above problem and increase the chance of detecting pitch accents without degrading the accuracy, we augment with a post-process based on heuristics – if an IP has no detected pitch accent, we will take the syllable with the highest accented score, i.e. the difference between the posterior probability of the accented syllables' model and the unaccented syllables' model, as a pitch accent. Furthermore, if the IP has many syllables (e.g. 7 or more), we take the syllable with the second highest accented score as a pitch accent as well.

With this post-process, the result is 72.86% in accuracy, as shown in TABLE VII, and its average EER for training data across the 10-fold cross-validation is 33.00%. The accuracy and EER are almost unchanged, but the percentage of accented syllables being detected as accented is improved from 17.04% to 53.02%.

TABLE VII. Pitch Accent Detection Results

Anno. / Rec.	H*	L*	Unaccented	?
Accented	1791	405	2103	176
Unaccented	1356	590	8676	268

Note: the cases of "*?" are omitted in calculating the accuracy.

B. *Performance of the intonation detector*

In this section, we evaluate the performance of the intonation detectors. To have a closer look at the importance of each of the selected features, we also evaluate the intonation detection performance using different combinations of the selected features, followed by a combination of all of them.

1) Performance of individual features: TABLE VIII shows that the performances of f_2 and f_3 are almost the same, for their values are close. The performance of f_1, whose location is the farthest one from the end of the IP, is the worst.

TABLE VIII. Detection Results Using Single Feature

Feature	f_1		f_2		f_3	
Anno. / Rec.	R	F	R	F	R	F
Rising	656	546	794	68	790	68
Falling	214	248	76	726	80	726
Avg. EER	39.29%		8.51%		8.72%	
Accuracy	54.32%		91.35%		91.11%	

Note: Avg. EER is the average EER for training data across the 10-fold cross-validation.

TABLE IX. Intonation Detection Results

Combination	f_1, f_2		f_1, f_3		f_2, f_3		f_1, f_2, f_3	
Anno. / Rec.	R	F	R	F	R	F	R	F
Rising	805	78	805	78	794	69	808	85
Falling	65	716	65	716	76	725	62	709
Avg. EER	8.54%		8.54%		8.48%		8.60%	
Accuracy	91.41%		91.41%		91.29%		91.17%	

Note: Avg. EER is the average EER for training data across the 10-fold cross-validation.

TABLE IX shows that the detectors with different feature combinations have similar performance as the detector using the single feature of f_2 or f_3. One reason is that the single features of f_2 or f_3 perform very well already, unless there are errors made by the pitch detector, which cannot be improved by the combination of the features. Another reason may be that

[1] EER is the point where the False Acceptance Rate (FAR) equals the False Rejection Rate (FRR) for a detection process. In our pitch accent detector, FAR is the percentage of all known unaccented syllables being falsely "accepted" as accented and FRR is the percentage of accented syllables being detected as unaccented ("rejected").

the number of IPs in our dataset (i.e. 1680) is still too small for training and testing the Gaussian discriminator.

Fig. 11 shows that the Gaussian discriminator can generally correctly classify the IPs' intonation as rising or falling. When pitch detector fail to detect pitch, or the pitch in a syllable seems irregular, e.g. with very low energy and very high pitch comparing to the neighboring values, the pitch will be set to -4, which is the reason many points fall on the line of -4 in Y (f_3) value.

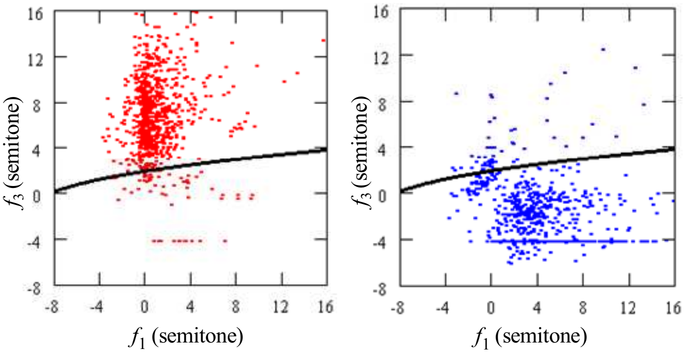


Figure 11. Distribution for the rising IP (the left figure) and the falling IP (the right figure), the line is the Gaussian discriminator

2) Analysis of the intonation detector using all 3 features: The detailed evaluation results of the detector using all of the 3 features are shown in TABLE X. It shows that 92.87% 808 / (62 + 808) of the annotated rising intonation can be correctly identified as rising intonation, and 89.29% 709 / (85 + 709) of the annotated falling intonation can be correctly identified as falling intonation.

TABLE X. INTONATION DETECTION RESULTS BASED ON THE MERGED DATA FOR TRAINING THE RISING AND FALLING GAUSSIAN DISCRIMINATORS.

Anno. **Rec.**	**R**	**U**	**L**	**F**	**?**
Rising	**808** (merged R, U)		**85** (merged L, F)		8
	774	34	3	82	
Falling	**62** (merged R, U)		**709** (merged L, F)		8
	49	13	5	704	

TABLE XI is the intonation detection results, which use the annotated pitch accents, instead of the recognized pitch accents. It shows that the accuracy has increased from 91.17% to 92.43%, and the average EER for training data across the 10-fold cross-validation decreases from 8.60% to 6.91%. It shows that there is still some room for improving the pitch accent detector to obtain further improvement in the overall intonation detection task.

TABLE XI. INTONATION DETECTION RESULTS USING THE ANNOTATED PITCH ACCENTS INSTEAD OF THE RECOGNIZED PITCH ACCENTS

Anno. **Rec.**	**R**	**U**	**L**	**F**	**?**
Rising	**823** (merged R, U)		**79** (merged L, F)		10
	788	35	3	76	
Falling	**47** (merged R, U)		**715** (merged L, F)		6
	35	12	5	710	
avg. EER: 6.91%			**Accuracy: 92.43%**		

Note: Avg. EER is the average EER for training data across the 10-fold cross-validation.

IX. CONCLUSIONS

We evaluated the intonation detection performance for ESL learners native in Mandarin. We have collected speech data from 40 ESL learners and performed cross-validation on the detection performance of the pitch accent detector and the intonation detector. We first develop a pitch accent detector based on the energy, pitch contour and duration of the vowels, and then choose 3 features based on the normalized values derived from the pitch of the nuclear tone. The pitch accent detector can correctly identify the syllables as accented or not in an accuracy of 72.86%, when its EER is 33.00%. The resulting intonation detector built for classifying ESL speech as either rising or falling intonation achieved a detection accuracy of 91.17%, with average EER of 8.60%.

ACKNOWLEDGEMENTS

This work is affiliated with the CUHK MoE-Microsoft Key Laboratory of Human-centric Computing and Interface Technologies. This projected is funded by the Research Grant Council, Hong Kong SAR government (CUHK 416108).

REFERENCES

[1] J. Anderson-Hsieh, R. Johnson and K. Koehler, "The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentals, prosody and syllable structure", Language Learning, vol 42, no. 4, pp. 529-555, 1992.

[2] S. Zhang, K. Li, W-K Lo, and H. Meng, "Perception of English suprasegmental features by non-native Chinese learners", Proc. the 5th Int. Conf. on Speech Prosody, Chicago, USA, May 2010, in press.

[3] H. Fujisaki, and K. Hirose, "Modelling the dynamic characteristics of voice fundamental frequency with application to analysis and synthesis of intonation", in Preprints of the Working Group on Intonation, The 13th Congress of Linguistics., Tokyo, pp. 57–70, 1982.

[4] D. Hirst, "Prediction of prosody: An overview", in Bailey, G. and Benoit, C. [Eds.], Talking Machines, North Holland, pp.199-204, 1992.

[5] P. Taylor, "The rise/fall/connection model of intonation". Speech Communication, vol. 15, pp.169-186, 1995.

[6] P. Taylor, "The Tilt intonation model", Proc. ICSLP 98, pp. 1383-1386, 1998.

[7] D. R. Ladd, Intonational Phonology (2nd ed), Cambridge University Press, pp. 87-107, 2008.

[8] J. B. Pierrehumbert. The phonology and phonetics of English intonation. MIT dissertation. 1980.

[9] A. Cruttenden, Intonation, Cambridge University Press, pp 57, 1986.

[10] A. Brugos, S. Shattuck-Hufnagel, and N. Veilleux. "Transcribing prosodic structure of spoken utterances with ToBI", MIT Open Course Ware. Online: http://anita.simmons.edu/~tobi/tutorial.html, accessed on 19 Mar, 2010.

[11] M. E. Beckman and G. Ayers-Elam, Guidelines for ToBI labelling, version 3.0. ms., Ohio State University. Online: http://www.ling.ohio-state.edu/~tobi/ame_tobi/, accessed on Mar 25, 2010.

[12] M. Selting, Prosodie im Gespräch. Max Niemeyer Verlag, 1995.

[13] F. Tamburini , "Prosodic prominence detection in speech", Proc. ISSPA, Paris, France, vol. 1, pp. 385 – 388, Jul. 2003.

[14] D. Wang and S. Narayanan, "An Acoustic Measure for Word Prominence in Spontaneous Speech", IEEE Trans. Speech and Audio Proc., vol. 15(2), pp. 690-701, 2007.

[15] K. Sjölander, "The Snack Sound Toolkit", Department of Speech, Music and Hearing, KTH Royal Institute of Technology, Online: http://www.speech.kth.se/snack/, accessed on Apr 29, 2010.

[16] S. Nooteboom. "The prosody of speech: melody and rhythm", in Hardcastle, Laver [Ed], The Handbook of Phonetic Sciences, pp. 640-673, Oxford: Blackwell Publishers, 1997.