



Using melody metrics to compare English speech read by native speakers and by L2 Chinese speakers from Shanghai.

Daniel Hirst¹, Hongwei Ding^{2,3}

¹Aix-Marseille University, LPL UMR 7309, 13100, Aix-en-Provence, France

²School of Foreign Languages, Shanghai Jiao Tong University, China

³Institute of Acoustics and Speech Communication, TU Dresden, Germany

daniel.hirst@lpl-aix.fr, hwding@sjtu.edu.cn

Abstract

In this study we analyse 18 metrics which were extracted fully automatically from the acoustic signal to describe the melodic characteristics of recordings of English read by L2 Chinese speakers from Shanghai. The metrics were compared to those of native English speakers recording the same material and also to comparable Chinese recordings read by the same speakers from Shanghai and also by other speakers. For the great majority of the metrics the values obtained for the L2 speakers were intermediate between those obtained from the recordings of English and of Chinese by native speakers.

Index Terms: L2 evaluation, speech prosody, melody metrics, Mandarin Chinese, English

1. Introduction

It is well established that the prosody of non-native speech is one of the major factors contributing to the perception of a foreign accent [1] and affecting the overall intelligibility of L2 speakers [2].

This paper is part of a larger project studying the prosodic parameters which could contribute to the automatic evaluation of the prosody of L2 speakers in general, and in particular of Chinese L2 speakers of English. More specifically, the project targets the production of Chinese speakers from Shanghai.

Among the prosodic parameters which can characterise a non-native speaker's production, rhythm is one of the most robust and well-studied (cf. [3] and [4]). [5] showed that rhythm metrics performed quite well in the automatic classification of speech samples: those produced by native (UK) and those produced by non-native (French) speakers of English, including two levels (intermediate and advanced) of non-native speakers.

As suggested by [6], the analysis of rhythm in speech and language should, perhaps, be extended to more than just the duration of segments or syllables. The recurring phonetic patterns in the production of speech form the characteristic of rhythm. The phonetic variables include not only patterns of syllabic timing, but also patterns of fundamental frequency and energy.

Speech melody is, however, notoriously more difficult to describe and to evaluate than the timing of speech segments, despite the fact that, intuitively, the melody of an utterance is one of its most perceptually salient characteristics.

A recent study [7] showed that a set of objective melody metrics performed quite well in distinguishing recordings of English, French and Chinese utterances read by native speakers of those languages. In this paper we apply a similar set of melody metrics to recordings of English spoken by Chinese speakers from Shanghai and compare them to the same metrics

obtained from recordings of English spoken by native speakers of English. We also compare the metrics to those obtained from the same Shanghai-Chinese speakers reading comparable material in Mandarin Chinese, as well as that from other speakers of standard Mandarin Chinese.

2. Data

2.1. Texts and recordings

Following [7], we used the texts from the Eurom1 corpus [8] which consist of 40 continuous, thematically connected passages, each of five sentences. The passages were originally composed in the 1980s and recordings made for 11 European languages. Recently, new recordings of these passages were made for English and French and the passages were translated, adapted and recorded for Korean and Mandarin Chinese, as part of the Open Multilingual Prosody Database (OMProDat) [9].

Unlike the original Eurom1 corpus, for which each speaker read only ten or fifteen passages, in the OMProDat recordings, for each language, ten speakers (5 male and 5 female) read all 40 passages. New recordings were then made of English and Chinese read by ten Chinese speakers (5 male, 5 female) from Shanghai, as part of a study to investigate dialect-specific characteristics of Shanghai speech that may influence the prosody of speakers in their L2 speech. Each speaker recorded the 40 passages first in their native language, then in English.

2.2. Melody metrics

Following the procedure used in [7], the fundamental frequency curves obtained from the recordings were processed using the Momel algorithm [10], implemented as a plugin to the Praat software [11], so that they could be represented as a sequence of points corresponding to the anchor-points¹ of a quadratic spline function, used to model the underlying pitch curve of the recording. This representation by a smooth continuous curve abstracts away from the microprosodic effect of the individual speech sounds. In order to eliminate a few erratic values, in this study we eliminated any anchor points which were less than 150 ms from the previous anchor-point. The anchor-points were then normalised to the Octave-median (OMe) scale [12], in order to eliminate speaker-specific differences, in particular differences between male and female speakers. The *ome* values were obtained using the formula:

¹In previous publications the term *target points* has been used to refer to these points. The term *anchor-points* is perhaps preferable to avoid confusion with the idea that these points are in any sense cognitive targets.

$$ome(f_0) = \log_2\left(\frac{f_0}{median}\right) \quad (1)$$

where *median* corresponds to the median value of f_0 for the whole five-sentence passage.

The following parameters were then calculated from the sequence of normalised anchor-points for each passage and each speaker:

pitch-m, pitch-sd mean and standard deviation of the OMe values of all anchor-points

high-m, high-sd mean and standard deviation of the OMe values of all anchor-points which are higher than the preceding point

low-m, low-sd mean and standard deviation of the OMe values of all anchor-points which are lower than the preceding point

interval-m, interval-sd mean and standard deviation of the octave differences between each anchor-point and the preceding point

rise-m, rise-sd mean and standard deviation of the differences between each anchor-point and the preceding point when the difference is positive

fall-m, fall-sd mean and standard deviation of the differences between each anchor-point and the preceding point when the difference is negative

slope-m, slope-sd mean and standard deviation of the intervals between consecutive anchor-points divided by the time intervals between the two points

rise-slope-m, rise-slope-sd mean and standard deviation of the intervals between consecutive anchor-points, when the interval is positive, divided by the time intervals between the two points

fall-slope-m, fall-slope-sd mean and standard deviation of the intervals between consecutive anchor-points, when the interval is negative, divided by the time intervals between the two points

These metrics were identical to those analysed in [7], except for the second and third pairs (*high* and *low*), which we added in this study to make the analyses more symmetrical. The metrics were calculated using a Praat script: *calculate_melody_metrics.praat*, which is available as an accompanying file to this text.

To illustrate from a hypothetical example, Figure 1 shows a quadratic spline curve defined by the five anchor-points: p1, p2, p3, p4 and p5. The values of *pitch* would be calculated as the mean and standard deviation of the OMe scale value (using equation 1) of points p1 to p5. The values of *high* would use only the values of p2 and p5 since these are the only values higher than the preceding point. Similarly the value of *low* would use the values p3 and p4 which are the only values which are lower than the preceding point. The values of *interval* and *slope* would be based on the pairs p1:p2, p2:p3, p3:p4 and p4:p5. Those of *rise* and *rise-slope* would use the values p1:p2 and p4:p5 while those of *fall* and *fall-slope* would use the values p2:p3 and p3:p4.

The parameters were calculated for each passage, for each of the ten speakers for Chinese spoken by native speakers (CNM²), Chinese spoken by native speakers from Shanghai

²This is the ISO 639-3 language code for Mandarin Chinese cf: http://en.wikipedia.org/wiki/ISO_639-3

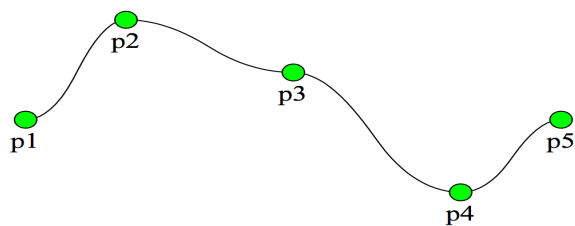


Figure 1: An example of a quadratic spline curve defined by 5 anchor-points.

(SH-CNM), L2 English spoken by Chinese speakers from Shanghai (SH-ENG) and English spoken by native speakers from the UK (ENG). The resulting data file thus contained a total of 1600 (=40*10*4) data points.

3. Analyses

The data obtained was analysed using the MASS package [13] of the R software [14]. Five analyses were carried out, each time comparing the metrics for two languages or varieties:

- English (ENG) vs. Standard Chinese (CNM)
- English (ENG) vs. Shanghai Chinese (SH-CNM)³
- Chinese (CNM) vs. Shanghai Chinese (SH-CNM)
- English (ENG) vs. Shanghai English (SH-ENG)
- Shanghai English (SH-ENG) vs. Shanghai Chinese (SH-CNM)

It is, of course, the fourth pair that particularly concerns us in this paper, since it provides us with a direct comparison of L1 speakers and L2 speakers reading the same texts.

We added the first pair to replicate the results of [7] with our new set of data and to provide a base-line for comparison with the other analyses. The differences in this replication were that we analysed recordings of 5 male and 5 female native speakers of English (compared to 5 male and 6 female speakers in [7]), and, as mentioned above, we eliminated anchor-points that were less than 150 ms from the preceding point, and we also added the four new metrics *high-m*, *high-sd*, *low-m*, *low-sd*.

We included the analysis of the second pair as a check that the melodic differences between English and Shanghai-Chinese were equivalent to those between English and standard Chinese.

We included the analysis of the third pair to quantify the differences between standard Chinese and Shanghai-Chinese.

We added the last pair to provide a direct comparison of the same speakers reading Mandarin Chinese and English.

For each comparison, we followed the same procedure. We first ran a principal component analysis, with the options of centring and scaling the data, using the function:

$$prcomp(data, center = T, scale = T) \quad (2)$$

The output of the *pca* analysis was submitted to a linear discriminant analysis. The significance of each individual metric for the fourth analysis was then tested using *anova*.

³Note that by *Shanghai Chinese* we mean *Mandarin Chinese* spoken by speakers from Shanghai, not the *Wu* dialect which is spoken in Shanghai and sometimes referred to as *Shanghai dialect* or *Shanghaihinese*.

4. Results

4.1. Analysis 1: Chinese (CNM) vs. English (ENG)

The linear discriminant analysis based on the Chinese and English recordings resulted in a confusion matrix as in table 1:

Table 1: Confusion matrix for the linear discriminant analysis of the melody metrics from the recordings CNM vs. ENG.

<i>predicted:</i>	CNM		ENG	
	F	M	F	M
CNM-F	132	50	11	7
CNM-M	60	133	0	7
ENG-F	7	1	118	74
ENG-M	7	1	37	155

This corresponds to a correct prediction of 94.88% of the language of the passages, which is close to perfect, especially considering that the melody metrics are obtained entirely automatically from the acoustic signal without reference to linguistic levels of representation such as phoneme, syllable or word. If we include the sex of the speaker in the analysis, then the prediction score drops to 67.25%. This is still much higher than the 25% we could expect from prior probabilities, but it also shows that the normalisation which we carried out on the data had the desired effect of reducing the differences between male and female speakers.

4.2. Analysis 2: Shanghai Chinese (SH-CNM) vs. English (ENG)

The comparison of the Chinese recordings by the speakers from Shanghai (SH-CNM) to the English recordings (ENG) by native speakers resulted in a confusion matrix as in table 2:

Table 2: Confusion matrix for the linear discriminant analysis of the melody metrics from the recordings SH-CNM vs. ENG.

<i>predicted:</i>	SH-CNM		ENG	
	F	M	F	M
SH-CNM-F	148	43	2	7
SH-CNM-M	53	138	3	6
ENG-F	2	5	120	73
ENG-M	6	2	37	155

This table show a correct prediction of 95.88% of the language of the passages, even higher than the score for English vs. standard Chinese. Once again the prediction including the sex of the speaker drops, this time to 70.13%, showing once again that the differences between male and female speakers was effectively to some extent neutralised.

4.3. Analysis 3: Chinese (CNM) vs Shanghai Chinese (SH-CNM)

Although this was not the central topic of this paper, it is interesting to see that the standard Chinese recordings and those by the speakers from Shanghai were reasonably well distinguished by the discriminant analysis which achieved a score of 63% correct prediction from the metrics analysed: still much higher than the 25% expected from prior probabilities but much lower than the other analyses. This score also further dropped (to 48.25%) when the sex of the speaker was included in the prediction.

Table 3: Confusion matrix for the linear discriminant analysis of the melody metrics from the recordings CNM vs. SH-CNM.

<i>predicted:</i>	CNM		SH-CNM	
	F	M	F	M
CNM-F	97	25	44	34
CNM-M	24	74	52	50
SH-CNM-F	30	11	128	31
SH-CNM-M	32	43	38	87

4.4. Analysis 4: English (ENG) vs Shanghai English (SH-ENG)

The fourth analysis concerns the central topic of this paper: the comparison between English read by native (UK) speakers, and English read by L2 speakers of Chinese from Shanghai. Here the discrimination reaches 78.75% from the metrics analysed. Once again the score dropped (to 57%) when the sex of the speaker was included in the prediction.

Table 4: Confusion matrix for the linear discriminant analysis of the melody metrics from the recordings ENG vs. SH-ENG.

<i>predicted:</i>	ENG		SH-ENG	
	F	M	F	M
ENG-F	117	40	18	8
ENG-M	40	101	38	21
SH-ENG-F	18	27	113	42
SH-ENG-M	8	28	39	125

Since this analysis was our principal focus here, we carried out an ANOVA for each of the 18 metrics described in 2.2.

Table (5) shows the significance level for each of the 18 metrics analysed for the factors *language* (L) and *sex* (S) and for the interaction between *language* and *sex* (LS).

Table 5: Significance levels of Anova for each metric. [-]: n.s., [*] = $p < 0.05$, [**] = $p < 0.01$, [***] $p < 0.001$

	mean			standard deviation		
	L	S	LS	L	S	LS
pitch	***	***	***	***	***	**
high	***	**	***	***	***	***
low	***	-	**	*	-	***
interval	***	-	-	-	***	***
rise	-	*	***	*	***	***
fall	-	**	***	-	***	***
slope	-	-	*	-	***	***
rise-slope	-	***	***	***	***	***
fall-slope	-	**	***	-	***	***

Space prohibits a detailed examination of these results but a regular pattern emerged from the comparison of the significant effects. We note, in particular, that the simple effects of *language* and *sex* were, for most of the metrics, far less significant than the interaction between *language* and *sex*. To take an example, the values for *rise-m* and *fall-m* showed a highly significant interaction, whereas the factor *language* was not significant for either metric and the factor *sex* was less significant for both metrics than the interaction between the two factors.

Figure 2 shows the distribution of mean values for rising intervals for male and female native speakers of English (ENG)

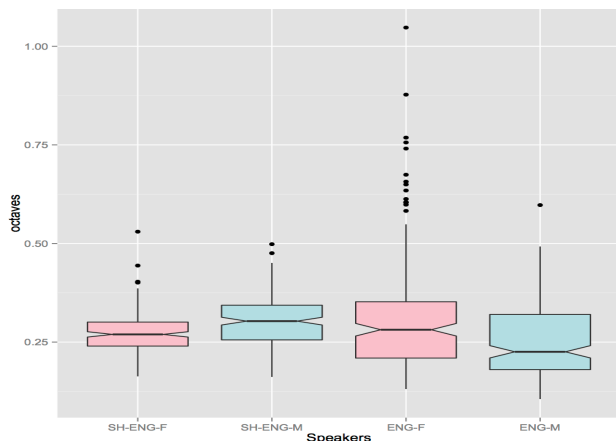


Figure 2: Distribution of mean values for rising intervals for male and female native speakers of English (ENG) and L2 speakers from Shanghai.

and L2 speakers from Shanghai (SH-ENG). As can be seen, there was a highly significant difference between the values for the male speakers, the L2 speakers having a much higher mean interval for the rises than the native speakers. The difference between the native and L2 female speakers, however was not significant, but there was a significant difference between male and female speakers of each group: female speakers having a smaller mean interval than the male speakers for the L2 speakers, while for the native speakers the mean interval was significantly larger for the female speakers than for the male speakers.

A similar effect is found for the metric *fall-m*, shown in Figure 3 except of course for the fact that the scale is reversed, since the intervals measured here were all negative. If we look at the absolute value of these intervals, however, we see that once again there is no significant difference between the native and non-native female speakers but a very significant difference between the native and non-native male speakers. And once again we find larger intervals for the native female speakers than for the native male speakers but smaller intervals for the non-native female speakers than for the native female speakers. A similar effect is found for practically all of the metrics examined in this study.

Interestingly, this is precisely the effect which was observed in [7], where most of the metrics analysed showed higher values for female speakers for the native English and French speakers but the opposite for the native Chinese speakers. We return to this in more detail in the final section of this paper.

4.5. Analysis 5: Shanghai English (SH-ENG) vs Shanghai Chinese (SH-CNM)

The final analysis compared the English readings of the Shanghai speakers to that of Chinese read by the same speakers. Here the correct language identification was even higher than in analysis 1 and analysis 2, reaching 96.38% correct identification. The correct identification of *language* and *sex* was 72%

5. Conclusions

The pattern which emerges from these analyses is that, for most of the metrics analysed, the non-native speakers showed values intermediate between those obtained for standard Chinese

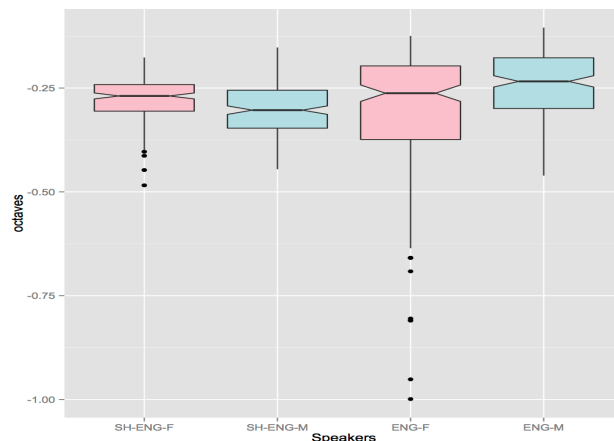


Figure 3: Distribution of mean values for falling intervals for male and female native speakers of English (ENG) and L2 speakers from Shanghai.

Table 6: Confusion matrix for the linear discriminant analysis of the melody metrics from the recordings SH-ENG vs. SH-CNM.

	SH-ENG		SH-CNM	
<i>predicted:</i>	F	M	F	M
SH-ENG-F	155	42	3	0
SH-ENG-M	59	134	3	4
SH-CNM-F	4	3	155	43
SH-CNM-M	2	10	59	137

(CNM) or Shanghai Chinese (CNM) and those for English spoken by native speakers. With the exception of the female native speakers of English, for 15 of the 18 metrics studied, (i.e. all except *pitch-m*, *interval-m* and *rise-slope-sd*), the order of the metrics was in each case CNM > SH-CNM > SH-ENG > ENG and in each case for the male speakers we found higher values for all the metrics than for the female speakers.

The exception to this was the female native speakers of English, whose metrics were all systematically higher than those of the male speakers. The reasons for this difference remain to be determined. [7] suggested that similar results which were obtained comparing recordings by English, French and Chinese native speakers could perhaps be explained by pressure from the constraints of producing lexical tone, which restricted the expressive use of pitch to express gender differences.

6. Acknowledgements

We thank Xiping Xu from Tongji University, Shanghai, for her help in the collection of the data for Shanghai subjects. The first author would like to acknowledge support from a 3-year contract (2011-2014) as lecture professor with Tongji University, Shanghai. The second author is sponsored by the National Science Foundation of China (13BYY009) and the Interdisciplinary Program of Shanghai Jiao Tong University (14JCZ03) for this research work.

7. References

- [1] M. Jilka, "The contribution of intonation to the perception of foreign accent." Ph.D. dissertation, Institut für Maschinelle Sprachverarbeitung, University of Stuttgart, 2000.
- [2] L. Hahn, "Primary stress and intelligibility: research to motivate teaching of suprasegmentals." *TESOL Quarterly*, vol. 38, pp. 201–223, 2004.
- [3] L. White and L. Mattys, "Calibrating rhythm: first language and second language studies." *Journal of Phonetics*, vol. 35, pp. 501–522, 2007.
- [4] H. Ding and R. Hoffmann, "A durational study of German speech rhythm by Chinese learners." in *Proceedings of the 7th International Conference on Speech Prosody (SP7)*, Dublin, Ireland, 2014, pp. 295–299.
- [5] A. Tortel and D. J. Hirst, "Rhythm metrics and the production of English L1/L2." in *Proceedings of the 5th International Conference on Speech Prosody 2010*, Chicago, USA., 2010, pp. P1–42.
- [6] K. J. Kohler, "Rhythm in speech and language: A new research paradigm," *Phonetica*, vol. 66, pp. 29–45, 2008.
- [7] D. J. Hirst, "Melody metrics for prosodic typology: comparing English, French and Chinese." in *Proceedings of Interspeech.*, Lyon, August 2013.
- [8] D. Chan, A. Fourcin, D. Gibbon, B. Granstrom, M. Huckvale, G. Kokkinakis, K. Kvale, L. Lamel, B. Lindberg, A. Moreno, J. Mouropoulos, F. Senia, I. Trancoso, C. Veld, and J. Zeiliger, "Eurom - a spoken language resource for the eu." in *Eurospeech'95. Proceedings of the 4th European Conference on Speech Communication and Speech Technology.*, vol. 1, Madrid, Spain., 18-21 September 1995, pp. 867–870.
- [9] D. J. Hirst, B. Bigi, H.-S. Cho, H. Ding, S. Herment, and T. Wang, "Building omprodat, an open multilingual prosodic database." in *Proceedings of TRASP, Tools and Resources for the Analysis of Speech Prosody [satellite workshop of Interspeech]*, Aix-en-Provence, August 2013, pp. 11–14.
- [10] D. J. Hirst, "A Praat plugin for Momel and INTSINT with improved algorithms for modelling and coding intonation," in *Proceedings of the XVIth International Conference of Phonetic Sciences*, Saarbrücken, 2007, pp. 1233–1236.
- [11] P. Boersma and D. Weenink, "Praat, a system for doing phonetics by computer." <http://www.praat.org> [version 5.4.06, February 2015], 1992 (2015).
- [12] C. De Looze and D. J. Hirst, "The OMe (Octave-Median) scale: a natural scale for speech prosody." in *Proceedings of the 7th International Conference on Speech Prosody (SP7)*, N. Campbell, D. Gibbon, and D. J. Hirst, Eds., Trinity College, Dublin, Ireland, May 2014.
- [13] W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S.*, 4th ed. New York: Springer, 2002.
- [14] R Core Team, *R: A language and environment for statistical computing.*, R Foundation for Statistical Computing, Vienna, Austria, 2012. [Online]. Available: <http://www.R-project.org>