Melody metrics for prosodic typology: comparing English, French and Chinese.

Daniel Hirst[1,2]

[1]Laboratoire Parole et Langage, UMR 7309 CNRS, Aix-Marseille University, France
[2]School of Foreign Languages, Tongji University, Shanghai, China
daniel.hirst@lpl-aix.fr

Abstract

In recent years there have been a number of proposals for objective paradigms for establishing prosodic typologies among languages. This paper compares the results of melody metrics calculated on just over two hours of read speech for each of three languages. Pitch movements in Chinese, a lexical tone language, were found to be significantly more ample and more varied than those obtained for English, characterised as a language with lexical stress and French, characterised as a language with no lexical prosody. Moreover, a gender difference, observed in both English and French was not observed in Chinese. It is conjectured that the pressure from the lexical use of tone in Chinese may have the effect of restricting the use of pitch for other functions.

Index Terms: speech prosody, melody, prosodic typology, objective metrics, automatic analysis.

1. Introduction

The study of the ways in which prosodic parameters, i.e. the relative length, pitch and loudness of individual speech sounds, are used in different languages is crucial to obtain a better understanding of the way in which speech prosody functions in human speech. Discovering reliable and robust objective *metrics* corresponding to these parameters would be extremely useful for improving our understanding of prosodic structure. These would also be likely to be useful in evaluating *atypical speech* such as non-standard dialects, non-native speech, pathological speech and synthetic speech. This, in turn, could be expected to contribute to improving current speech synthesis systems and to guiding automatic speech recognition.

2. Prosodic typology

Linguists have proposed a number of prosodic typologies to describe the variability across languages, although the relationship between these typologies and objective measurements of acoustic signals has often proved elusive.

Among these, the most well-established is the *lexical prosodic typology*, distinguishing: **quantity languages** with lexically distinctive long and short sounds (e.g. *Korean* and *Finnish*) **tone languages**, with lexically distinctive high and low pitch (e.g. *Chinese* and *Hausa*) and **stress languages**, with lexically distinctive prominent and non-prominent syllables (e.g. *Arabic* and *English*).

Also very commonly used in linguistic descriptions, is the *rhythmic typology* [1, 2, 3] classifying languages as **stress-timed**, e.g. *English*, *Russian*, *Arabic*, **syllable-timed** e.g. *French*, *Telugu*, *Yoruba* and **mora-timed** e.g. *Japanese*, *Tamil*, *Ganda*.

There have also been a few proposals for a *melodic typology* based on the recurrent pitch patterns commonly found in languages. The comparison of English and French, two of the languages analysed in this paper, [4, 5] brought to light a distinction between the underlying pitch patterns of the two languages. Abstracting away from more global intonation patterns, accented words in English, on these analyses, are basically associated with an underlying **falling pitch pattern**, whereas they are associated with a **rising pitch pattern** in French. This phonological characterisation, however, undergoes a number of local modifications, so that the actual observed surface configurations may be quite different from these more abstract underlying patterns.

3. Prosodic metrics

The proposed prosodic typologies described in the preceding section have led phoneticians to search for corresponding metrics: objective measurements which would make it possible to predict the typological category to which the language belongs from the acoustic data.

3.1. Lexical prosodic metrics

The automatic classification of languages according to the lexical typology described above does not seem to have been the subject of systematic research (but see [6] on rhythm), although a number of linguists have pointed out the difficulties of deciding, for example, whether a particular prosodic system is basically tonal or accentual. This remains very much an area for future research.

3.2. Rhythm metrics

After a number of years of disappointing results in establishing objective criteria for the classification of languages on the basis of their rhythmic properties [7, 8], the last decade has seen a revival of interest in defining robust metrics which correlate with the rhythmic typological distinctions. Among these are **[%V]**: the percent duration of vocalic intervals in an utterance [9], **[ΔC, ΔV]**: the standard deviation of the duration of consonantal and vocalic intervals [9], **[rPVI (c,v)]** and **[nPVI (c,v)]**: the raw and normalised indices of variability between duration of successive consonantal and vocalic intervals and **[VarcoC,VarcoV]**: the coefficient of variation of duration of consonantal and vocalic intervals [10].

It was shown [11] that a linear discriminant analysis using a combination of these metrics provided an efficient discrimination of native speakers of English and intermediate French learners of English reading the same texts. Results for advanced French learners of English were intermediate between those of the other two groups.

3.3. Melody metrics

Surprisingly, there has been less research into the area of *melody metrics*, despite the fact that work on automatic language identification [12] showed that including prosodic parameters, led to an improvement in the performance of a segmental based language identification system when applied to four languages (English, Spanish, Japanese and Chinese) chosen as representatives of different typological groups. Overall features derived from measurements of pitch were found to be the most useful for discrimination.

One study [13] showed that the analysis of pitch points calculated using the Momel algorithm (see below section (5)) reveals significant effects for language and gender of speakers for five languages (English, French, German, Italian, Spanish).

In [14], I measured mean, standard deviation and coefficient of variation for pitch intervals and slopes of rises and falls in continuous passages in French and English. These metrics were also calculated from consecutive target points derived automatically from the recordings using the Momel algorithm described below section (5)).

The identification of the language attained 87.6% correct discrimination from these metrics, which were based solely on the distribution of the pitch target-points, without any consideration of the relationship of the target points to phonological, lexical or syntactic constituents. As noted at the time [14], the small number of recordings analysed, 150 five-sentence passages for English and 100 for French, makes it difficult to generalise about prosodic differences between the two languages.

4. Building OMProDat: an open multilingual prosodic database

In order to provide a firmer basis for the analysis of prosodic metrics, our laboratory decided to build an open multilingual prosodic database **OMProDat**)[15], to be archived and distributed by the recently created *Speech and Language Data Repository (SLDR)* (http://sldr.org) under an open database license. The aim of this database is to collect, archive and distribute recordings and annotations of directly comparable data from a representative sample of different languages representing different prosodic typological characteristics.

Two of the studies reported in the previous section [13, 14] both used the 40 five-sentence continuous passages taken from the Eurom1 corpus, recorded as a deliverable of the European SAM (Speech Assessments and Methodology) project [16]. The passages were originally recorded in the 1980's by ten speakers (five male and five female), but each speaker read only a limited number of the 40 passages (15 for English and 10 for French). In addition, the original recordings of the corpus were under copyright, owned by the different laboratories and universities which had produced the recordings.

We consequently decided to make new recordings of this corpus, with all 40 passages read by 10 speakers, and which would be distributed by the *SLDR* as part of the *OMProDat* database, under an open-database license.

The first language recorded under these conditions was Korean [17]. This was followed by new recordings for English and French read by native speakers, as well as English read by native speakers of French and French read by native speakers of English [18]. Most recently we have added recordings for Standard Chinese [19].

It is intended that all these corpora shall be annotated using the automatic annotation tools described in the next section, and that all the recordings and annotations will be made freely available under an open-database licence as part of the open multilingual speech-prosody database. Linguists and engineers are welcome to make use of the corpora and are asked, in return, to make any additional annotations which they may carry out publicly available on the *SLDR*.

5. Automatic prosodic annotation

5.1. Automatic phonetisation and alignment

The analysis of the prosodic structure of speech requires the alignment of the speech recording with a phonetic transcription of the speech, usually in some version of the International Phonetic Alphabet. This task is extremely labour-intensive - it may require several hours for even an experienced phonetician to transcribe and align a single minute of speech manually. It is consequently obvious that transcribing and aligning several hours of speech by hand is not generally something which can be envisaged.

A number of tools can currently be used to automate the task, including the *HTK Toolkit* [20], *Festival* [21], *Julius* [22], the *P2FA* [23], and *EasyAlign* [24].

SPPAS (SPeech Phonetisation Alignment and Syllabification) [25] is a recently developed GPL licensed set of tools for Unix based platforms, including Linux, Mac-OSX and Windows (using Cygwin).

The tool implements the automatic phonetisation and alignment of speech from an orthographic transcription of the recording and is specifically designed to be used by linguists in conjunction with other tools for the automatic analysis of speech prosody. It is currently implemented for French, English, Italian and Chinese and there is a simple procedure to add other languages.

SPPAS generates separate TextGrid files for *inter-pausal units*, *words*, *syllables* and *phonemes*. The output is illustrated as a Screenshot in Figure (1), with TextGrid files merged in Praat [26]. For this example in Chinese, only 3 tiers are illustrated, with the 'tokens' tier corresponding to both *syllables* and *morphemes*, since the corpus was transcribed in Pinyin.

5.2. Automatic annotation of pitch

There are currently a number of different algorithms for the automatic annotation of pitch contours.

The Momel algorithm [27, 28] assumes that a pitch contour can be adequately represented by a sequence of target points, each contiguous pair of which is linked by a continuous smooth monotonic quadratic interpolation (defining a quadratic spline function). This, in turn, assumes that the shape of a pitch contour is entirely determined by the temporal and frequential values of the relevant target points. The algorithm uses a form of robust regression to optimise the modeling of raw fundamental frequency curves with a quadratic spline function.

The Momel algorithm has been implemented as a Praat plugin [27], making it possible for a linguist to use its functions directly from the Praat menus without needing to handle scripts. An evaluation of the improved algorithm carried out on a corpus of read speech in Korean [29] showed a significant and systematic improvement as compared to the older version of the algorithm.

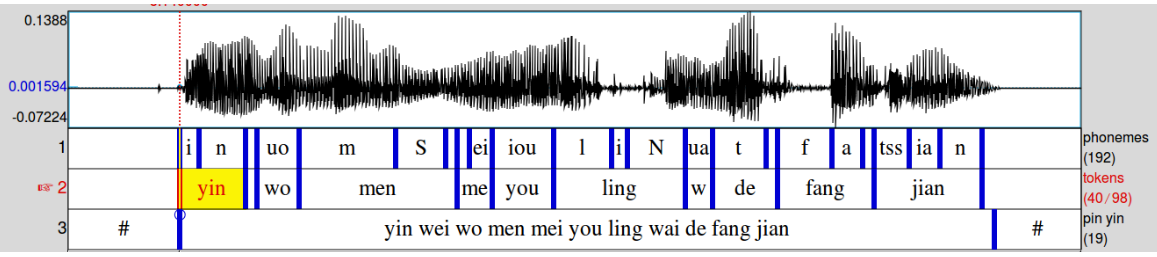
0.1388
0.001594
-0.07224
1
2
3
i n uo m S ei iou l i N ua t f a tss ia n
yin wo men me you ling w de fang jian
# yin wei wo men mei you ling wai de fang jian #
phonemes (192)
tokens (40/98)
pin yin (19)

Figure 1: SPPAS *output example for the utterance in Chinese (Pinyin)* Yinwei women meiyou lingwai de fangjian *(Because we do not have another room)*

Since it was first developed, the Momel algorithm has been applied to a number of different languages, including English, French, Italian, Catalan, Brazilian Portuguese, Venezuelan Spanish, Russian, Arabic, isiZulu and Korean (for references see [27]). More recently ([30]), the algorithm was applied to a corpus of speech in Standard (Beijing) Chinese. This was particularly challenging, since the corpus used was spontaneous speech and involved a language with a rich lexical tone system.

6. Melody metrics revisited

6.1. Data and metrics

For this study the English (EN), French (FR) and Chinese (ZH) recordings from the *OMProDat* database were annotated using *SPPAS* and the pitch contours were modelled as target points using *Momel* as described above. These consisted of 40 five-sentence passages each read by at least ten speakers for each language (EN: 5m, 6f; FR: 4m, 7f; ZH: 5m, 5f), a total of 2000 or 2200 sentences per language, or just over two hours of recordings.

In order to reduce the inter-subject variability, the target points were scaled using the OMe (Octave-Median) scale proposed in [31] with the formula:

$$ome = log_2(Hz/median) \qquad (1)$$

where *median*, here, is the median value of f_0 for the whole five-sentence passage.

From the scaled target points the mean and standard deviation for each passage was calculated for:

octave value of target points on the OMe scale

interval absolute difference from previous target

rise, fall difference from previous target for rise and fall separately

slope absolute difference from previous target divided by distance in seconds

rise-slope, fall-slope slope for rise and fall separately

In [14] the coefficient of variation was used as well as the mean and standard deviation since the metrics were calculated on the raw f0 values. Here this is not necessary since all the values were offset to the speaker's median f_0 by (1) so 14 values were calculated for each passage.

6.2. Results

The data was subjected to a linear discriminant analysis resulting in the confusion matrix of Table (1).

	Predicted		
	English	*French*	*Chinese*
English	**339**	84	17
French	148	**241**	48
Chinese	17	52	**328**

Table 1: Confusion matrix for discriminant analysis on language.

We can see from this that the discrimination of English from French with these 14 parameters is just over 71%. This is lower than in the earlier study but here the discrimination is based on 440 passages per language rather than 100 and 150. The global discrimination of the three languages is also just over 71%. The discrimination of Chinese as against French and English, however, is much higher, reaching 89%.

If we include gender as a factor in the analysis, the global discrimination rises to 74% correct prediction of both language and gender as shown in the corresponding confusion matrix in Table (2).

	Predicted					
	EN-f	*EN-m*	*FR-f*	*FR-m*	*ZH-f*	*ZH-m*
EN-f	**186**	0	44	5	1	4
EN-m	0	**172**	0	22	0	6
FR-f	49	0	**202**	0	27	1
FR-m	9	87	26	**34**	1	1
ZH-f	3	0	33	0	**164**	0
ZH-m	1	2	0	11	0	**183**

Table 2: Classification matrix for discriminant analysis on language and gender.

Note, as we saw above, that the data here has been offset to the speaker's median pitch so there is no simple factor of pitch height that is used in this analysis. The prediction of language regardless of gender is also slightly better in this analysis (76%), while the discrimination of English and French from Chinese now reaches 93%.

Table (3) shows the significance level for each of the 14 parameters analysed for *language* (L) and *gender* (G) and for the interaction *language*gender* (L*G).

Space prohibits a detailed analysis of these results here but as an example the box-plot in Figure (2) shows that *pitch rises* are, on average, of greater amplitude for Chinese than for French, and for French than for English.

For English and French there is a significant gender difference: for both languages the pitch rises of female speakers

	mean			standard deviation		
	L	G	L*G	L	G	L*G
octave	***	-	***	***	***	***
interval	***	-	*	***	***	***
rise	***	***	***	***	***	***
fall	***	***	***	***	***	***
slope	-	-	-	***	-	-
rise-slope	***	***	***	***	-	-
fall-slope	***	***	***	***	-	-

Table 3: Significance levels of Anova for each parameter. [-] : n.s., [*] = p¡ 0.05, [**] = p ¡ 0.01, [***] p ¡ 0.001

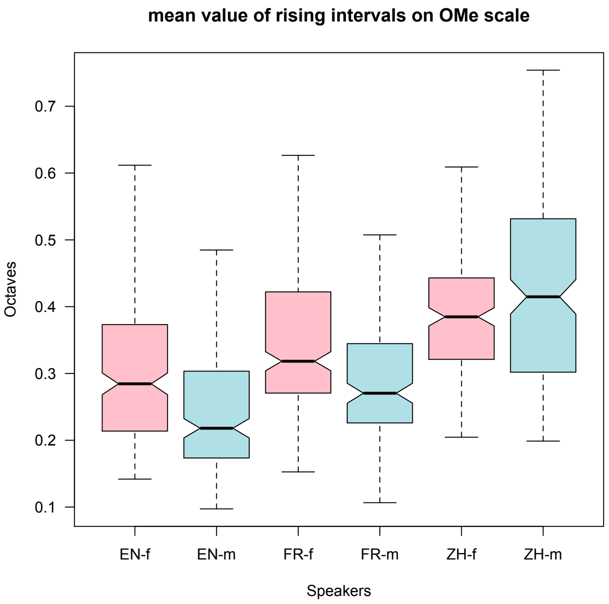

Figure 2: Mean value of rising intervals using the Octave-Median scale.

are significantly greater than for male speakers. Note, once again, that, since all the values are calculated on the octave-median scale, the greater amplitude of the pitch rises for female speakers is not simply an effect of higher pitch. For Chinese, the gender difference is reversed: male Chinese speakers produced larger pitch rises than female speakers but the difference is probably not significant as can be seen from the fact that the "notches" on the boxes overlap for these two.

Interestingly, a similar pattern is repeated for all the parameters for which the ANOVA reveals a significant difference: in each case we find a highly significant difference between Chinese on the one hand and English and French on the other hand. Between English and French we find a smaller and, for some parameters, non-significant difference. For English and French we find a significant gender difference whereas for Chinese the gender difference is reversed but generally non-significant.

These results show that in Chinese, pitch movements are larger (mean *interval*, *fall* and *rise*), with greater variability (sd of *interval*, *fall* and *rise*) and also faster (mean *slope*, *rise-slope*, *fall-slope*) than in English and French contrary to previous results [32].

Furthermore, in English and French there is a very significant gender difference (female speakers make larger and faster pitch movements) which is not observed in Chinese - if anything the tendency is reversed: male speakers making larger and faster pitch movements.

These preliminary results suggest the possibility that the lack of gender difference in Chinese could be the result of pressure from the lexical tone, that limits the use of pitch in Chinese for such non-lexical functions.

Considerably more data from these and other languages needs to be examined, however, before this interpretation can be seriously entertained.

7. Conclusion and perspectives

The examination of melody metrics measured on a total of a little more than six hours of read speech produced by 10 English, 10 French and 10 Chinese speakers reveals a significant difference between Chinese on the one hand and English and French on the other. For the two European languages there was also a significant gender difference which was not observed for Chinese.

It would be particularly interesting as a next step to extend the paradigm to other languages both with and without lexical tone. The results presented here are based solely on the linear sequence of the target points as obtained by the Momel algorithm, without any consideration of the relationship of the target-points to phonological, lexical or syntactic constituents.

The automatic alignment of the recordings using the SPPAS algorithm will allow us to extend these analyses to take into consideration the position of the pitch movements with respect to these different constituents. We hope to present results from this in future work.

8. Acknowledgements

Most of the data presented here was presented orally in [33] but there was no corresponding written presentation.

Thanks to my colleagues Brigitte Bigi and Sophie Herment from the *LPL*, CNRS, Aix-en-Provence, and Hongwei Ding and Ting Wang from the *SFL*, Tongji University, Shanghai, as well as to my former doctoral students Céline De Looze, Anne Tortel and Hyongsil Cho, and to Anastassia Loukina and Greg Kochanski from the *Oxford University Phonetics Laboratory*, all of whom helped with different aspects of data collection and analysis for this project.

9. References

[1] Abercrombie, D. "Syllable quantity and enclitics in English", in In Honour of Daniel Jones. Papers contributed on the occasion of his eightieth birthday, 12 September 1961., D. Abercrombie, D.B. Fry, P.A.D. MacCarthy, N.C. Scott, and J.L.M. Trim, (eds.). Longmans, London.: 216-222. 1964.

[2] Pike, K.L. The Intonation of American English. The University of Michigan Press, Ann Arbor. 1945.

[3] Ladefoged, P. A Course in Phonetics. New York: Harcourt College Publishers, 2001.

[4] Hirst D.J.; Di Cristo, A. "A survey of intonation systems", In D.J. Hirst and A. Di Cristo, editors, Intonation Systems: A Survey of Twenty Languages, chapter 1, pages 144. Cambridge University Press, 1998.

[5] Jun, S.-A.; Fougeron, C. "A phonological model of French intonation", in Intonation: Analysis, modelling and technology, 209242, 2000.

[6] Farinas, J.; Pellegrino, F. “Automatic rhythm modelling for language identification”, Proceedings of Eurospeech 2001, 2539-2542. 2001.

[7] Roach, P. “On the distinction between stress-timed and syllable-timed languages”, In D. Crystal, ed. Linguistic Controversies. Edward Arnold, London, 1982.

[8] Bertinetto, P. M. “Reflections on the dichotomy stress vs. syllable-timing”, Revue de Phonétique Appliquée, 919293:99130., 1989.

[9] Ramus, F.; Nespor, M.; Mehler, J. “Acoustic correlates of linguistic rhythm in the speech signal”, Cognition, 73:265292, 1999.

[10] Dellwo, V.; Wagner, P. “Relations between language rhythm and speech rate”, Proceedings of the 15th international congress of phonetics sciences, pages 471474, 2003.

[11] Tortel, A.; Hirst, D.J. “Rhythm metrics and the production of English L1/L2”, In Proceedings of the 5th International Conference on Speech Prosody 2010, pages P142, Chicago, USA., 2010.

[12] Thymé-Gobbel, A.; Hutchins, S. “On using prosodic cues in automatic language identification”, In Proceedings ICSLP 96, 17681771, Phliadelphia, 1996.

[13] Campione, E; Véronis, J. “A statistical study of pitch target points in five languages”, In Proceedings of ICSLP, Sydney, 1998.

[14] Hirst, D.J., “Pitch parameters for prosodic typology. A preliminary comparison of English and French”, In Proceedings of the XVth International Congress of Phonetic Sciences, Barcelona, 2003.

[15] Hirst, D.J.; Bigi, B.; Cho, H.-S.; Ding, H.; Herment, S.; Wang, T. “Building OMProDat, an open multilingual prosodic database”, Proceedings of TRASP, Tools and Resources for the Analysis of Speech Prosody, satellite workshop of Interspeech 2013, Aix-en-Provence, August 30, 2013.

[16] Chan, D. Fourcin, A.; Gibbon, D.; Granstrom, B.; Huckvale, M.; Kokkinakis, G.; Kvale, K.; Lamel, L.; Lindberg, B.; Moreno, A.; Mouropoulos, J.; Senia, F.; Trancoso, I.; Veld, C.; Zeiliger, J. Eurom - a spoken language resource for the EU. In Eurospeech95. Proceedings of the 4th European Conference on Speech Communication and Speech Technology., 1, 867870, Madrid., 18-21 September 1995.

[17] Kim, S.H. Hirst, D.J.; Cho, H.-S.; Lee, H.-Y.; M. Chung, M.H. Korean Multext: A Korean prosody corpus. In Proceedings of the 4th International Conference on Speech Prosody., Campinas, Brazil., 2008.

[18] Herment, S.; Loukina, A;. Tortel, A.; Hirst, D.J.; Bigi, B. “Aix-ox: A multi-layered learners corpus: automatic annotation. 4th International Conference on CorpusLinguistics., Jaèn, Spain, 2012.

[19] Ding, D.; Hirst, D.J. ‘A preliminary investigation of third-tone sandhi in Standard Chinese with a prosodic corpus”, 8th International Symposium on Chinese Spoken Language Processing, Hong Kong 2012.

[20] Young, S.Y.; Young, S. The HTK hidden markov model toolkit: Design and philosophy. vol. 2, pp. 244, Entropic Cambridge Research Laboratory, Ltd, 1994.

[21] Taylor, P. A.; Black, A.; Caley, R. “The architecture of the Festival speech synthesis system”, In Proceedings of the Third ESCA Workshop in Speech Synthesis, 147151, Jenolan Caves, Australia, 1998.

[22] Lee, A; Kawahara, T.; Shikano, K. “Julius — an open source real-time large vocabulary recognition engine”, In Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH). 1691–1694, 2001.

[23] Yuan, J.; Liberman, M. Speaker identification on the scotus corpus. In Proceedings of Acoustics 2008, 56875690, 2008.

[24] Goldman, J.-P. “EasyAlign: a friendly automatic phonetic alignment tool under Praat”, In Proceedings of Interspeech 11., n: Ses1-S3:2, Florence, Italy, 2011.

[25] Bigi and D. J. Hirst. “SPeech Phonetization Alignment and Syllabification (SPPAS): a tool for the automatic analysis of speech prosody”, In Proceedings of the 6th International Conference on Speech Prosody., May 2012.

[26] Boersma, P.; D. Weenink, D. Praat, a system for doing phonetics by computer. http://www.praat.org [version 5.3.41, February 2013], 1992 (2013).

[27] Hirst, D.J. “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: 12331236, Saarbrucken, 2007.

[28] Hirst, D.J. “The analysis by synthesis of speech melody: from data to models”, Journal of Speech Sciences, 1(1): 5583, 2011.

[29] Hirst, D.J.; Cho, H.S.; Kim, S.H; Yu, H. “Evaluating two versions of the momel pitch modeling algorithm on a corpus of read speech in korean”, In Proceedings of Interspeech 8, 16491652, Antwerp, Belgium, September 6-10 2007.

[30] Zhi, N.; Hirst, D.J.;Bertinetto, P.M. “Automatic analysis of the intonation of a tone language. applying the momel algorithm to spontaneous Standard Chinese (Beijing). In Proceedings of Interspeech 11, Makuhari, Japan, September 26-30 2010.

[31] De Looze, C.; Hirst, D.J. “L’echelle OME (Octave-MEdiane): une echelle naturelle pour la melodie de la parole.” in Actes des XXVIIIes Journes d’Etude sur la Parole, Mons, Belgium, May 25-28 2010.

[32] Xu, Y.; Sun, X. “Maximum speed of pitch change and how it may relate to speech”, Journal of the Acoustical Society of America, 111:13991413, 2002.

[33] Hirst, D.J., “The automatic analysis by synthesis of Speech Prosody with Preliminary Results on Mandarin Chinese”, 8th International Symposium on Chinese Spoken Language Processing, Hong Kong, [Invited keynote lecture]. 2012.