Automatic analysis of the intonation of a tone language. Applying the Momel algorithm to spontaneous Standard Chinese (Beijing).

Na Zhi[1], *Daniel Hirst*[2], *Pier Marco Bertinetto*[1]

[1]Laboratorio di Linguistica, Scuola Normale Superiore, Pisa, Italy
[2]Laboratoire Parole et Langage, CNRS & Université de Provence, France

na.zhi@sns.it, daniel.hirst@lpl-aix.fr, pier.marco@bertinetto.eu

Abstract

This paper describes the application of the Momel algorithm to a corpus of spontaneous speech in Standard (Beijing) Chinese. A selection of utterances by four speakers was analysed automatically and the resynthesised utterances were evaluated subjectively with two categories of errors: lexical tone errors and intonation errors. The target points determining the pitch contours of the synthetic utterances were then corrected manually in order to obtain a set of acceptable utterances for the entire corpus. An application attempting to optimise window-size for the Momel algorithm showed no overall improvement with respect to the manually corrected data. This annotated data will nevertheless constitute a useful yardstick for evaluating improvements to the automatic algorithm which is expected to be far more robust than data annotated for languages with no lexical tone.

Index Terms: intonation, modelling, spontaneous speech, Chinese, Momel

1. Introduction

The study reported in this paper is part of a larger, ongoing project concerned with the description of the prosody of spontaneous speech in various languages, building on results obtained in earlier work carried out at the *Scuola Normale Superiore* in Pisa, Italy [3] and in the CNRS *Laboratoire Parole et Langage* in Aix-en-Provence, France [8, 9]. In our general approach we take the view that a satisfactory model of speech should meet the epistemological requirements of explicitness, predictivity and unification ([3]). One way to meet at least the first two of these requirements, is by a paradigm of *analysis by synthesis* [10, 12]. If the representation which is derived from the analysis can be used to provide a satisfactory synthetic output, with no significant loss of information, then we can conclude that this representation contains the relevant information that we are hoping to capture. Synthesis, in this perspective, is seen as a tool for the evaluation of the quality of analysis rather than as an end to itself.

In order to achieve really adequate descriptive models of the prosody of different languages, it will be necessary to go far beyond the often fairly limited data samples which are generally used to construct such models. This will obviously imply treating large quantities of data for which it will not be feasible to envisage entirely manual annotation. Techniques of automatic analysis are consequently particularly attractive for this type of study.

In this paper we are particularly concerned with the description of the intonation of spontaneous speech in Standard Chinese (Beijing). In the field of intonation analysis, one explicit modelling approach is the Momel algorithm [7, 9] (see 3.2), which has so far been successfully used for the modelling of the intonation of a number of different languages (for references see [13]). This study is, however, the first time that the algorithm has been systematically applied to the intonation of a language with a complex tonal system like Chinese, and in particular to a corpus of spontaneous speech in such a language.

We present here the first results of an ongoing evaluation of potential problems in the application of the Momel algorithm to a corpus of spontaneous speech in Standard Chinese. A selection of utterances was analysed automatically and then subjected to manual corrections. It is hoped that the resulting database with hand-corrected Momel annotation will be a useful tool for the development of improved automatic algorithms which will be essential for the development of phonological models on the basis of large-scale corpora.

2. Tone and intonation in Standard Chinese

Modelling the intonation of Chinese is particularly challenging because of the intertwined relationship between Chinese lexical tones and intonation in a raw F0 contour.

2.1. The intertwining of tone and intonation

According to [1], the communicative use of sentence intonation in tone languages is just as free as in non-tone languages. The study of intonation in a tone language, however, is greatly complicated by the multiple uses of pitch for both lexical tone and intonation. These two phonological strands are:

> phonetically intertwined in the tempo and pitch contour of an utterance [2]

which is quite unlike the F0 pattern of a completely non-tonal language like French, Finnish, or Korean, where all the pitch events of F0 contour are determined by the supra-lexical functions of intonation. In other languages, like English, Russian, Arabic, the existence of lexical accent, or word-stress, does contribute to some extent to the overall makeup of a global pitch patten, but the individual lexical items do not have predefined pitch configurations associated with them in the way that languages like Chinese, Yoruba or Vietnamese do.

In Chinese, although tone and intonation are independent as linguistic functions and phonological categories, the phonetic output of the two are inseparable from each other in an F0 contour. The interactive relation between lexical tones and the purely intonational use of tune [16] make the whole F0 picture intricate, and in need of further exploration.

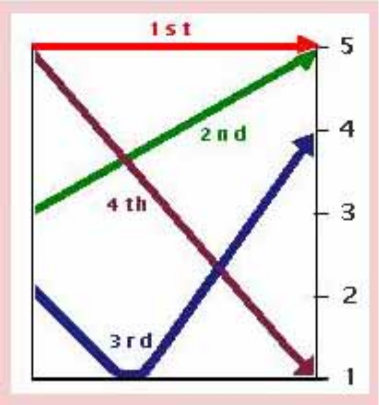

Figure 1: *Schematic diagram of canonical Standard Chinese lexical tones* (from Chao [5]).

2.2. Lexical tone in Chinese

In the historical literature, the superimposed relation of Chinese lexical tones and intonation tunes is described vividly in a metaphor by Chao ([6] as, small ripples riding on top of large waves, where lexical tones are compared to small ripples in an ocean, which add on the large waves of the overall intonation tune, resulting in an algebraic sum of the two waves.

In Standard Chinese (Beijing) there exist four canonical lexical tones, which are marked in written form with an iconic diacritic, above the nucleus of the associated syllable, such as, zhī (meaning to know), zhí (straight), zhǐ (paper), zhì (intelligent). The four tones are also categorized numerically as T1, T2, T3 and T4, so the above four tonal syllables can also be annotated as: zhi1, zhi2, zhi3 and zhi4. For the phonetic representation of the four tones, Chao [5] proposed a numerical notation system with five scaling points, 1 to 5 corresponding respectively to low, mid-low, mid, mid-high and high position within a speakers normal pitch range. Therefore, based on native speakers acoustic perception, the four lexical tones can be represented with a succession of numerals marking respectively the significant points of each tonal contour. Thus, T1 is transcribed as [55], T2 as [35], T3 as [214], and T4 as [51], indicating the high level form of T1, the high-rising (from mid to high) form of T2, the low-falling-rising form of T3 and the high falling form of T4. The canonical form of the four tonal shapes and heights (in relative values) can be seen in Fig. 1.

In the phonetic analysis of intonation with spontaneous data, there is a big challenge due to the unpredicted manifestation of lexical tones in an F0 contour. In the speech flow, the actual phonetic values of tones are only rarely close to their underlying pitch configurations, and they present far more variable features than might be expected. Some common tonal changes found in speech can be explained by *sandhi* rules (see [6]: 27-28). However, when we look at the actual f0 data of fluent running speech, various allotonal phenomena in Chinese are beyond the interpretive capacity of phonological rules. The extensive undershoot of tones is often too distorted to be easily recognized, and a limited number of tone sandhi rules are far from being sufficient to account for all such variation.

The actual surface manifestation of tones is greatly influenced by a number of factors such as the neighbouring tonal context, the modified pitch range, the speed of utterance, the sentence focus, the speakers emotion etc. From our data, it can be seen that a global pattern of intonation may have little resemblance to the concatenation features of the consituant tones. In continuous speech, an underlying rising form of a T2 may actually show a falling pattern, and there are many other examples involving such reversed tonal contours in speech. Some syllables may lose their pitch features partially or even entirely.

Tseng ([18]) also observed in her spontaneous data that there is often no match between the phonological prediction and the final phonetic output of lexical tones. However, she noted that although a substantial portion of a sentence might be produced with incomplete acoustic information, listeners have no difficulty in understanding the sentence meaning, for they can integrate the phonetic input with their native knowledge of the language.

According to Shen [17], there are five components contributing to the final phonetic output of speech:

a lexical tones at the basic morphemic level,

b normal stress denoting possible lexical stress at the boundary of word or phrase level,

c tune at the sentence level expressing sentence modality

d emphatic focus

e emotive intonation at the sentence level superimposed on the overall pitch movement of the utterance.

3. Applying the Momel algorithm to a corpus of spontaneous Chinese

3.1. Corpus

The corpus employed in our study consists of conversational recordings taken from the Chinese Spontaneous Conversation Corpus (or CADCC) produced by the Chinese Academy of Social Sciences, Beijing [15, 14]. The corpus consists of 12 units of daily conversation with no specific topics defined between native Beijing speakers. Each unit has two different speakers involved and each dialogue lasts around one hour. In this study, 100 utterances were chosen from the CADCC corpus. Among them, 60 utterances (30 utterances from each speaker) are from a dialogue between two female speakers, and 40 utterances (20 utterances from each) are from a dialogue between two male speakers. The selected utterances are neutral in emotion and vary in sentence length and syntactic types, and include as few as possible disturbing speech variants such as overlaps, laughing, and background noise, etc. The interjections and pauses within the utterances were all labeled manually by the first author in the TextGrid tier, and annotated so that they would not be taken into account in the acoustic analysis.

3.2. The Momel algorithm

The Momel algorithm [7, 9] analyses a raw fundamental frequency curve as the product of two components: a global *macroprosodic* component, corresponding approximately to the underlying intonation pattern of the utterance, and a local *microprosodic* component, representing deviations from the macroprosodic curve which are caused by the articulatory constraints of individual phone segments. The discontinuity observed in a raw fundamental frequecy curve is thus modelled by the microprosodic component, which means that the underlying macroprosodic curve can be modelled as a continuous and smooth curve, such as is found when speech is ”re-iterated” by humming or by pronouncing entirely sonorant sequences like /mamama/ or /lalala/. In the Momel algorithm, this underlying

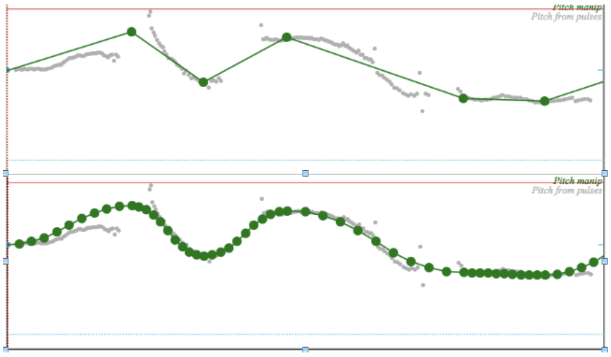

Figure 2: *Example of a pitch curve overlaid with the target points obtained from the Momel algorithm with, top panel: linear interpolation, bottom panel: quadratic spline interpolation.*

curve is modelled as a quadratic spline function, the simplest mathematical function which respects the constraints of continuity and smoothness while interpolating between a sequence of 'target points'. The algorithm takes as input a raw F0 curve and gives as output a corresponding sequence of target points for the quadratic interpolation. The algorithm has recently been implemented [11] as a *plugin* for the Praat speech analysis environment [4].

The Momel algorithm thus constitutes an example of the analysis by synthesis paradigm we mentioned in section 1. Once we obtain a satisfactory set of representations of this type we shall then be in a position to look into the relationship between these target points and more abstract higher-level representations such as that provided by the Intsint coding system, [9] or other more ToBI-like systems such as that described in [16]. For the purposes of this study we will restrict our attention to this first level of abstraction from the raw data in the belief that it will prove sufficiently theory-friendly (if not entirely theory-neutral) to be useful in different theoretical paradigms.

3.3. Analysing the data with Momel

A representative sample of 100 utterances produced by two male and two female speakers was selected for this study. The utterances were first coded using the automatic algorithm with default parameters [9]. The utterances were then re-synthesised using Psola resynthesis with the original f0 replaced by the quadratic spline curve interpolating between the target points output by the analysis.

In some cases the use of the quadratic interpolation was crucial to obtain the re-synthesis of the correct lexical tones. Figure 2 shows an example of the pitch curve of the utterance 'Na4 shi2 hou0 xia4 xiang1 de0 shi2...' In the top panel, the target points are shown connected by straight lines and in the bottom panel, the same points are shown connected by a quadratic spline function. When the utterance is synthesised with quadratic interpolation then the correct tones are perceived on all the syllables. When the utterance is synthesised with linear interpolation, however, then the high level tone of the fifth syllable 'xiang1' is perceived incorrectly as a falling tone.

Each of the utterances was consequently synthesized with the values from the automatic coding and quadratic spline interpolation. The synthetic utterances were then categorised by the first author using the following categories:

a. lexical tone error - one or more of the lexical items in the utterance was perceived as being pronounced with the wrong lexical tone.

b. intonation error - the resynthesised utterance was perceived as being produced with the correct lexical tones but with a perceptibly different intonational meaning.

c. correct tone and intonation - even if the utterance did not sound exactly the same as the original, there was no perceptible lexical or intonational difference of meaning.

Table 1 shows the number of syllables which were manually corrected from the total corpus, tabled by tone and by speaker.

Table 1: Number of syllables for each tone for each speaker in our corpus showing (in brackets) the number of these syllables which require manual correction

	speakers			
	A	B	C	D
T1	63 (12)	73 (4)	47 (3)	23 (0)
T2	83 (12)	86 (13)	59 (10)	36 (8)
T3	52 (14)	62 (8)	41 (9)	41 (9)
T4	107 (22)	144 (19)	85 (7)	63 (10)
T0	57 (7)	53 (6)	35 (2)	23 (0)

In the above table, *A*, *B*, *C* and *D* represent the four speakers of the study. *A* and *B* are female speakers, *C* and *D* male speakers. Horizontal lines refer to tone types (*T1*, *T2*, *T3*, *T4* plus *T0* for neutral tone). Digits refer to the number of syllables of the given tone type produced by the given speaker. Digits within brackets indicate the number of syllables that required manual correction, due to error in automatic pitch detection.

Utterances of categories a and b were then subjected to manual correction using the Praat plugin facility [11]. The position of the target points were adjusted manually until all the resynthesised utterances could be assigned to category c.

4. Using the database to optimise the modelling

As described above, the strategy of correction that was adopted was designed so that the output would be a set of Momel target points which correspond to acceptable equivalents to the original utterances.

As such, these target points can be thought of as a limiting set that we can hope to approach either by modifying the algorithm itself (cf [13]) or by adapting the parameters of the algorithm so that it is more suited to the specific needs of a particular corpus.

In the case of a tonal language like Chinese, it might be thought that tonal events, being partly determined by lexical constraints, are likely to be more frequent and rapidly executed than in languages with no tonal contrasts such as those to which the Momel algorithm has so far generally been applied. By contrast, Xu [19] has claimed that in both Mandarin Chinese and English, tonal patterns are produced at close to the maximum speed possible for pitch change.

4.1. Optimising window size for the Momel algorithm

If tonal events are more frequent in spoken Chinese than in English or French, we might expect that the default window size for the algorithm (300 ms) which was obtained using data from languages with no lexical tone would not be optimal for Chinese.

In order to test this we carried out automatic evaluation of the Momel target points with a variable window size ranging from one half of the default length to 1.5 times the default length by steps of 50 ms. The values tested, then, were 150, 200, 250, 300, 350, 400 and 450 ms.

For each of the seven values of window size, the complete set of 100 utterances was analysed. For each utterance, we calculated the output of a quadratic spline interpolation between the target points given by the algorithm. The correlation between this curve and that obtained from the manually corrected curve was calculated. Following [13] we decided to eliminate from the analysis those portions of the curve corresponding to potentially voiceless segments of the utterance. In fact, we decided to include in the correlation analysis only those segments of an utterance corresponding to the rime of a syllable. This is justifiable from the fact that in Standard Chinese, both the nucleus and the coda are always fully voiced whereas the onset may be voiced or voiceless.

The correlations reported here, then, correspond to the portions of the quadratic spline curves corresponding to the nucleus or the coda of a syllable. These portions were identified from manually labelled TextGrid files associated with each of the 100 utterances.

An analysis of variance of the correlations obtained for each of the 100 utterances factors showed no significant effect for window size (F(6, 700)= 0.732, p=0.619). We interpret this to mean that there is no systematic improvement to the algorithm to be obtained simply by reducing the size of the analysis window.

5. Conclusions and perspectives

A first attempt at applying the Momel algorithm to a corpus of spontaneous speech from Standard Chinese, a language with lexical tone, was relatively encouraging. Even though we were not able to make any systematic improvement to the Momel algorithm by globally modifying the analysis window size for the corpus that we analysed, we believe that the database and the evaluation technique described in this paper will be a particularly valuable tool for the evaluation of future adaptations of the algorithm, since the fact that the majority of errors corrected manually were lexical errors makes the data far more robust than data annotated for languages with no lexical tone.

6. Acknowledgements

The first author benefited from a PhD scholarship from the Scuola Normale Superiore da Pisa as well as financial support for a three month academic mission at the Laboratoire Parole et Langage in Aix-en-Provence.

7. References

[1] Abramson, Arthur S. and Svastikula, K. Intersections of tone and intonation in Thai. Reprint of the *3rd international Congress of Linguistics*. 13-22. 1982.

[2] Beckman, Mary E. *Stress and Non-stress Accent*. Dordrecht, Holland: Foris Publications. 1986.

[3] Bertinetto, Pier Marco and Bertini, Chiara. Towards a unified predictive model of Natural Language Rhythm. *Quaderni del Laboratorio di Linguistica della SNS* 2007-08.

[4] Boersma, Paul and Weenink, David. (1992-2010). *Praat: a system for doing phonetics by computer*. Available from http://www.praat.org.

[5] Chao, Yuen Ren. 1930. A system of tone-letters. *Le Matre Phontique* 45. 2427.

[6] Chao, Yuen Ren. 1933. Tone and intonation in Chinese. Bulletin of the Institute of History and Phiology 4. 121-134.

[7] Hirst, Daniel and Espesser, Robert . Automatic modelling of fundamental frequency using a quadratic spline function. *Travaux de l'Institut de Phontique d'Aix* 15. 71-85. 1993.

[8] Hirst, Daniel. Automatic analysis of prosody for multi-lingual speech corpora. In Keller E., G. Bailly, A. Monaghan, J. Terken & M. Huckvale (eds.) 2002. *Improvements in Speech Synthesis*. Chichester, England: John Wiley & Sons Ltd. 320-328. 2001.

[9] Hirst, Daniel. Form and function in the representation of speech prosody. in K.Hirose, D.J.Hirst, Y.Sagisaka (eds) *Quantitative prosody modeling for natural speech description and generation* (=*Speech Communication* 46 (3-4)), 334-347. 2005.

[10] Hirst, Daniel. A framework for the multilingual analysis by synthesis of speech melody. (Keynote lecture) *Proceedings Autumn conference, Korean Society for the Study of Speech Science and Technology*, 39-44. Seoul National University, novembre 17-18. 2006.

[11] Hirst, Daniel. A Praat plugin for MOMEL and INTSINT with improved algorithms for modelling and coding intonation. In *Proceedings of the 16th International Congress of Phonetic Sciences*. Saarbrcken. Germany. 1233-1236. 2007.

[12] Hirst, Daniel and Auran, Cyril. Analysis by synthesis of speech prosody: the ProZed environment. *Proceedings of Interspeech/Eurospeech 05. 9th European Conference on Speech Communication and Technology*, September 2005, Lisbon. 3225-3228. 2005.

[13] Hirst, Daniel; Cho, Hyongsil; Kim, Sunhee; Yu, Heun Evaluating two versions of the Momel pitch modelling algorithm on a corpus of read speech in Korean. *Proceedings of the VIIIth Interspeech Conference* (Antwerp 2007) pp. 1649-1652. 2007.

[14] Li, Aijun Chinese Prosody and Prosodic Labeling of Spontaneous Speech. *Proceedings of First International Conference on Speech Prosody* Aix-en-Provence, 2002.

[15] Li, Aijun; Zheng, Fang; Byrne, William; Fung, Pascale; Kamm, Terri, Ruhi, Umar, Venkataramani, Veera and Chen, Xiaoxia . CASS: A phonetically transcribed corpus of Mandarin spontaneous speech. *Proceedings of the VIth ICSLP Conference* Beijing, China. 2000.

[16] Peng, Shu H.; Chan, M.K.M.; Tseng, Chiu-Yu; Huang, Tsan; Lee, Ok J. and Beckman, Mary E. Towards a pan-mandarin system for prosodic transcription. In JUN Sun Ah (ed.) 2005. *Prosodic Typology: The Phonology of Intonation and Phrasing*. New York: Oxford University Press Inc. 230-270. 2005.

[17] Shen, Xiao N *The Prosody of Mandarin Chinese*. Berkeley, CA: University of California Press. 1990.

[18] Tseng, Chiu-Yu *An Acoustic Phonetic Study on Tones in Mandarin Chinese.* Brown University. Ph.D. dissertation [Institute of History & Philology, Academia Sinica] 1981.

[19] Xu, Yi and Sun, X. Maximum speed of pitch change and how it may relate to speech. *Journal of the Acoustical Society of America* 111. 1399-1413. 2002.