Can we model pitch using only the f0 on sonorant rimes?

Daniel J. Hirst, Ting Wang

Can we model pitch using only the f0 on sonorant rimes?

Daniel Hirst[1], *Ting Wang*[2]

[1]Laboratoire Parole et Langage UMR 7309
CNRS & Aix-Marseille University, Aix-en-Provence, France
[2]School of Foreign Languages, Tongji University, Shanghai, China
daniel.hirst@lpl-aix.fr, 2011ting_wang@tongji.edu.cn

Abstract

Modelling pitch patterns from acoustic data needs to take into account the fact that raw f0 curves are the product of an underlying global pitch pattern and a more local (micromelodic) influence of the individual speech sounds. This suggests the hypothesis that pitch could be modelled using only the f0 detected on sonorant rimes (vowels and sonorant codas). This paper describes an experiment to test the hypothesis. The test used recordings of Mandarin Chinese, assuming that evaluating synthetic prosody in a tone language would be a less metalinguistic task than in a language with no lexical tones. After applying an automatic alignment algorithm to the recordings, two resynthesized versions were created: in the first, only the f0 on sonorant rimes was used for the model. In the second the complete f0 curve was used. In both versions the f0 was modeled using the Momel algorithm. The recordings were then evaluated by 10 native speakers of Mandarin Chinese. Contrary to our hypothesis, the version using only the f0 detected on sonorant rimes was evaluated as significantly much worse than the standard method of using the whole f0 curve. A number of reasons for this difference are discussed.

1. Introduction

There have been many different approaches to modelling fundamental frequency patterns for speech synthesis or analysis. One approach, developed in particular by researchers from the "Dutch school" [1], was to develop a model of the way in which pitch is perceived. This was done by stylising raw fundamental frequency patterns as a sequence of straight lines, such that when the stylised frequency is used to re-synthesise the utterance, the result is judged to be perceptually equivalent to the original intonation pattern.

Another approach to modelling pitch has been to attempt to model the way in which pitch is produced by speakers. In particular, work by Fujisaki and his colleagues applied a model of pitch production [2] to a large number of languages, including several tone-languages, analysing an intonation pattern as the superposition of three underlying components: a global baseline component, a sequence of phrasal components and a sequence of shorter accent components. These three components are added in the logarithmic domain to produce a raw fundamental frequency curve.

A third approach has been to develop acoustic models which are neither directly models of speech perception nor of speech production but which are intended to be compatible with both.

Whatever the approach, fitting a raw f0 curve with a mathematical model is not a simple straightforward problem, due to the fact that fundamental frequency curves are not always continuous: unvoiced portions of the utterance have no associated f0. Even when the curve is continuous it is often not smooth and this type of irregularity can be very hard to model simply.

The discontinuity and irregularity of the f0 curve is generally due to the presence of obstruents in the utterance: stops and constrictives, which either interrupt the curve (voiceless obstruents) or make it irregular (voiced obstruents). The effect of these consonants has been called *micromelodic* as distinct from the *macromelodic* characteristics of larger pitch movements associated with accents and intonation patterns [3].

Micromelodic effects, then, are caused by the aerodynamic characteristics of the articulation of different phonemes. Phonemes like vowels and sonorants, which hardly obstruct the airflow, have virtually no micromelodic effect, whereas stops and constrictives interrupt or disturb the flow.

Linguists have know for a long time that fundamental frequency curves obtained from utterances containing only sonorants and vowels are much better behaved than raw f0 curves obtained from unrestricted speech. It is for this reason that linguists have often constructed sentences consisting of mainly sonorants and vowels such as Eva Gårding's *Madame Marianne Mallarmé har en mandolin från Madrid.* (Madam Marianne Mallarmé has a mandolin from Madrid) for Swedish [4], Annti Iivonen's *Laina lainaa Lainalla lainen.* (Laina lends Laina a loan) for Finnish [5] or Hiroya Fujisaki's *Aoi aoinoewa yamanouenoieni aru.* (The picture of the blue hollyhock is in a house on top of the hill) for Japanese [6].

2. Hypothesis

Since any intonation pattern can presumably be produced on any utterance, including on utterances with only voiceless obstruents, it seems reasonable to assume that only the pitch produced on the vowels (as well perhaps as on sonorant syllable codas, see below) is relevant for the perception of pitch patterns. This is in agreement with Nooteboom's observation [7] that we do not perceive the observable discontinuities of raw pitch-patterns unless they are longer than about 200 ms., as if human perception unconsciously bridges the silent gap by filling in the missing part of the pitch contour. A second, related, assumption, which also seems reasonable, is that speakers do not need to take into account the segmental makeup of the utterance when producing an intonation

pattern.

We can illustrate these two assumptions with two phrases in French: *à ma maman* (to my mummy), which contains only sonorant segments and *à ton papa* (to your daddy) in which all obstruents are voiceless. Figure (1) shows the signal and pitch for a recording of these two phrases pronounced with a declarative intonation pattern and Figure (2) shows the same phrases pronounced with an interrogative intonation pattern.

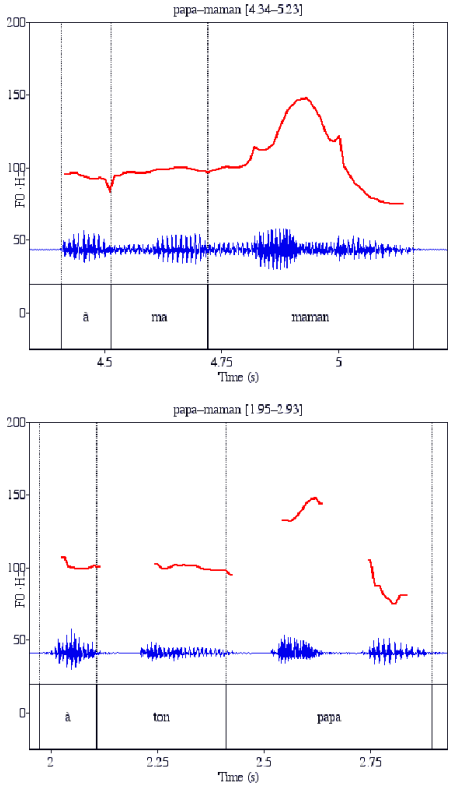

Figure 1: *Signal and pitch for the French utterances: 'A ma maman' and 'A ton papa' pronounced with a declarative intonation pattern.*

These assumptions seem to justify the hypothesis that pitch patterns could be modelled using only the information on the vowels (or sonorant rimes). As far as we know, this hypothesis has never been tested empirically.

The raw fundamental frequency curve can be thought of as the interaction between two orthogonal components: a micromelodic component, which is conditioned by the segmental nature of the individual speech sounds, and a macromelodic component, which corresponds to the underlying laryngeal gesture. This idea of factoring a raw fundamental frequency curve into two components is the rationale behind the Momel pitch modelling algorithm [8, 9].

The Momel algorithm was specifically designed to be used without needing any linguistic annotation of the f0 curve. When the program was first developed, software for the automatic alignment of phonetic transcriptions with the speech signal was far less readily available than it is today, so it made sense to try to model the f0 curve without using any information about the segmental nature of the utterance.

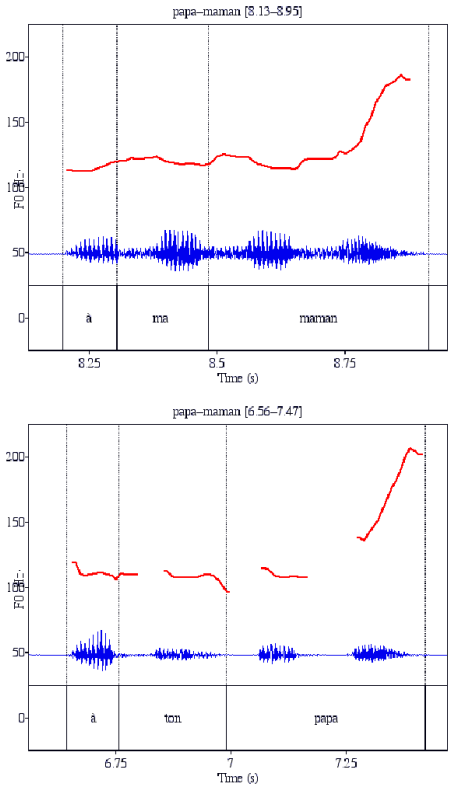

Figure 2: *Signal and pitch for the French utterances: 'A ma maman' and 'A ton papa' pronounced with an interrogative intonation pattern*

Today there exist a number of automatic alignment systems (such as SPPAS [15]) which make it possible to use the segmental information in the analysis of the pitch.

If the result of modelling the pitch using only the f0 on the sonorant rimes is as good as when using the whole f0 curve, this could make the task of the modelling algorithm much simpler, since we would no longer need to adapt the algorithm to the micromelodic characteristics. It would then justify rethinking the Momel algorithm significantly.

In the rest of this paper we present results from a preliminary experiment designed to compare the result of modelling intonation patterns using only the pitch detected on the sonorant rimes of syllables with the more usual practice of using the whole fundamental frequency curve for the model.

3. Method

We decided to compare the result of applying the Momel algorithm to the whole fundamental frequency curve with that of applying the algorithm to the curve where all the values except those of the sonorant rimes have been set to 0 (= unvoiced).

We were immediately faced with the problem of evaluating the modelled pitch patterns - it is notoriously difficult to elicit judgments of wellformedness of intonation patterns. One of the main reasons for this is that judging

wellformedness is a metalinguistic task, which is, necessarily, heavily dependent on the subjects' linguistic training and experience.

Pitch, is, however, used to distinguish lexical items in several languages, generally categorised as tone languages. The evaluation of the wellformedness of pitch patterns in a tone language is, consequently, at least partially, a genuine linguistic task, since the subject is asked to decide *what* was said rather than evaluating *how* it was said. Detecting errors of tone in a tone language are, thus, similar to detecting errors of phonemes in other languages.

Standard descriptions of modern Chinese describe the tones of Mandarin as high level, rising, falling rising and falling. The four words:

(1)

猫	毛	泖		帽
māo	máo	mǎo		mào
cat	*hair*	*still*	*water*	*cap*

all have different meanings as shown in the glosses.

In Mandarin Chinese [10, 11], all syllables end with a vowel, a semivowel or a final nasal sonorant (/n/ or /ŋ/); there are no final obstruents. For this reason, in our analysis we decided to take the sonorant rime, rather than just the vowel, as the potential carrier of the relevant pitch.

4. Material and subjects

For the test, we used 23 continuous passages, each of five sentences, read by one female native speaker of Mandarin Chinese (f01), taken from the OMProDat database OMProDat-cmn01 (Mandarin Chinese) [12, 13].

The pitch for each of the 23 recordings was modelled with the Momel algorithm under two conditions:

Version (A). The pitch was detected using the two-pass method described in [14]. The phonetic transcription of the text was aligned with the acoustic signal using the SPPAS software [15]. Pitch values for segments other than vowels and sonorant codas were then set to 0 (unvoiced), including 20 ms of the preceding and following vowel in order to eliminate the most significant micromelodic effects of the consonants on adjacent vowels.

Figure 3 shows an example of the resulting f0 curves used for the two conditions. The f0 used for version A has been lowered in the display (but not of course in the experiment) by 10% to improve legibility.

The pitch curves thus obtained were then modelled with the Momel algorithm and a synthetic version of each recording was created using the overlap and add algorithm in Praat, replacing the original pitch by the quadratic spline curve obtained from the anchor points with the Momel algorithm.

Version (B). The synthetic version was obtained using the same procedure as in version A, except that the Momel algorithm was applied directly to the pitch detected using the two-pass method.

We made two lists. List 1 contained a pseudo random alternation of version A and version B recordings of the 23 passages, as in:

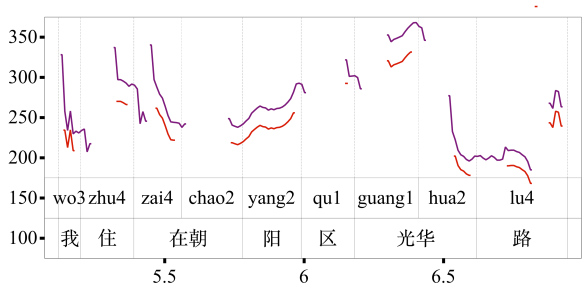

Figure 3: *An example, from sentence 2 of passage 10 of the f0 used for modelling pitch in our experiment. The lower curve (loweredin this display by 10% to improve legibility) was used for version A and the upper curve was used for version B.*

(2)

A	*B*	*A*	*A*	*B*	*A*	*B*	*B*	*A*	*B*	*A*	*A*	*B*	*A*	*B*	**...**
1	2	3	4	4	5	5	6	7	7	8	9	9	10	10	...

where the letter corresponds to the version and the number to the passage. The list was designed so that each time that two versions of a passage were presented, the A version was always presented before the B version.

List 2 was obtained by replacing the A version of the recording by the B version and vice versa so that in this list, the B version of a recording was always presented before the A version. Each list contained a total of 35 recordings.

The subjects who took part in the test were 10 native speakers of Mandarin Chinese: 5 male and 5 female students from Tongji University, aged between 23 and 33 years. The subjects were asked to listen to each passage from one of the lists and to highlight on a written copy of the text any characters (= syllables) that they felt had not been pronounced correctly. Since the only modification made to the original recordings was the f0 contour, we expected the subjects to interpret "correctly" as meaning with a different tone and hence potentially with a different meaning, even though the context is usually sufficient to interpret the meaning correctly, just as it is with errors of phonemes in languages without lexical tones.

The subjects were free to listen to each passage as many times as they liked but had to listen to.the whole passage without interruption each time.

Example (3) shows a sample annotation of version A of passage t01 where the subject has highlighted 10 characters as being incorrectly pronounced.

(3) 上星期我的一个朋友去医院打预防针她打算去中东度假为此她需要打一些预防霍乱伤寒甲肝小儿麻痹和破伤风的预防针我想打完所有这些针后她一定感觉很不舒服是她自己 决定要把这些针一次都打完的我可没办法同情她

The number of characters highlighted by each subject for each version of each passage was recorded and the mean number of errors for each version of each passage was used as the dependent variable for the experiment.

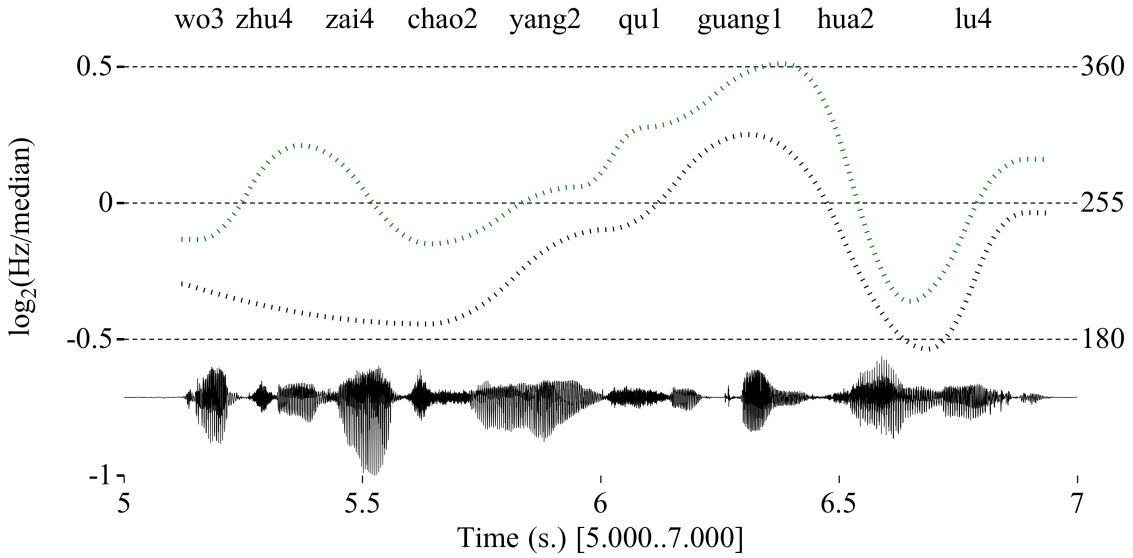

Figure 4: *Example of the pitch modelling with the two algorithms. (A) using only the pitch on sonorant rimes (lower curve) and (B) using the pitch on the whole utterance (upper curve). The display of the pitch of the A curve has been lowered by 10% to improve legibility*

5. Results

To test for a possible effect of order of presentation, we first calculated a paired t-test on the mean number of errors for each recording by list, since, as we noted above, in list 1 the A version always preceded the B version whereas in list 2 the B version always preceded the A version.

The effect of list was completely non-significant: t = -0.53345, df = 34, p-value = 0.5972.

We then tested the difference between the A version and the B version of the recordings. Here the difference was extremely significant: t = 5.5958, df = 34, p-value = 2.887e-06 with a mean difference of 2.68 between the two versions. As can be seen in Figure 5, the mean number of errors for the A version was significantly much higher than the mean number of errors for the B version.

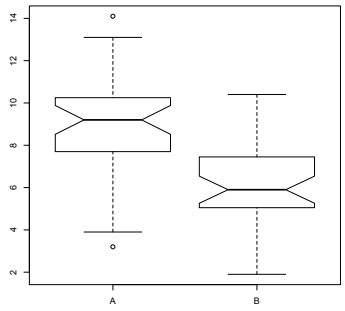

Figure 5: *Boxplot of mean number of errors by version: Version A using only the f0 on sonorant rimes and version B using the whole f0 curve.*

Figure 4 shows the modelled pitch curve for versions A and B obtained by applying the Momel algorithm to the f0 curves in Figure 3.

In version A of this sentence, the first syllable, *wo3*, sounds like tone 1 instead of tone 3, the second syllable *zhu4* sounds like tone 1 instead of tone 4. Syllable 6, *qu1* sounds too high, like a continuation rise after the preceding tone 2 syllable. Version B of this sentence sounds essentially correct.

An informal inspection of the data showed that many of the errors occurred on syllables with tone 3, which was not pronounced low enough, or on syllables with tone 2, which didn't rise high enough. Many of these syllables had a stop or affricate onset.

6. Conclusions and perspectives

Contrary to the hypothesis we mentioned in section (2), it appears that attempting to model the pitch of utterances using only the f0 observed on sonorant rimes in Mandarin Chinese, results in a speech synthesis which is evaluated as significantly much worse than when the complete f0 curve is used for the modelling.

There are a number of reasons which might contribute to this result.

The first could be an effect of the automatic alignment algorithm, which was not manually corrected in our test. This was a deliberate choice since we are interested in a fully automated technique which does not rely on manual correction, but it is true that an inaccurate alignment of the phonetic transcription could result in excluding crucial parts of the f0 curve.

An informal inspection of the automatic alignment, though, suggests that while there were some errors in the alignment, these would not be enough to explain the very significant difference in the evaluation of the two synthetic versions.

A second possible cause is the fact that we included

20ms of adjacent vowels when we set the pitch on obstruents to unvoiced. This was perhaps too drastic since some short vowels around 50 ms long, as in the syllable *qu1* of Figure 3, which also illustrates a case where the final boundary of the syllable is certainly placed too early.

Further testing will be necessary to see whether we can replicate the results reported here when devoicing *only* the f0 of obstruents. There are one or two other possible factors which we are currently investigating in more detail and which we hope to be able to report on during the presentation of our paper.

While it seems that it is not currently possible to use only the f0 on sonorant rimes to model pitch in our corpus, we now need to run further tests, taking into account all the factors we mentioned above.

7. References

[1] J. 't Hart, R. Collier, and A. Cohen. *A Perceptual Study of Intonation: an Experimental-Phonetic Approach to Speech Melody*. Cambridge University Press, 1990.

[2] H. Fujisaki. "Modeling the generation process of F0 contours as manifestation of linguistic and paralinguistic information". In *Proceedings of the XIIth International Congress of Phonetic Sciences*, pp. 1–10, 1991.

[3] A. Di Cristo and D.J. Hirst. "Modelling French micromelody: analysis and synthesis". *Phonetica*, 43(1- 3), pp. 11–30, 1986.

[4] E. Gårding. "Intonation in Swedish". In D.J. Hirst and A. Di Cristo, eds, *Intonation Systems. A Survey of Twenty Languages.*, chapter 6, pp. 117–136. Cambridge University Press, Cambridge, 1998.

[5] A. Iivonen. "Intonation in Finnish". In D.J. Hirst and A. Di Cristo, eds, *Intonation Systems. A Survey of Twenty Languages*, chapter 17, pp. 331–347. Cambridge University Press, 1998.

[6] H. Fujisaki. "Information, prosody and modeling - with emphasis on tonal features of speech". In *Proceedings of the 2nd International Conference on Speech Prosody*, pp. 1–10, 2004.

[7] S. Nooteboom. "The prosody of speech: melody and rhythm". In W. Hardcastle and J. Laver, eds, *The Handbook of Phonetic Sciences.*, pp. 640–673. Blackwell, Oxford, 1997.

[8] D.J. Hirst and R. Espesser. "Automatic modelling of fundamental frequency using a quadratic spline function". *Travaux de l'Institut de Phonétique d'Aix*, 15, pp. 75–85, Jan 1993.

[9] 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*, pp. 1233–1236, Saarbrucken, 2007.

[15] B. 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.

[10] P. Kratochvil. *The Chinese Language Today: Features of an Emerging Standard.* Hutchinson, London, 1968.

[11] S. Duanmu. *The Phonology of Standard Chinese*. Oxford University Press, 2nd. edition, 2007 (2000).

[12] H. Ding and D.J. Hirst. A preliminary investigation of the third tone sandhi in Standard Chinese with a prosodic corpus. volume Proceedings of the 8th International Symposium on Chinese Spoken Language Processing, Hong Kong, December 2012.

[13] 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], pages 11–14, Aix-en-Provence, August 2013.

[14] C. De Looze and D.J. Hirst. "The OMe (Octave-Median) scale: a natural scale for speech prosody". In N. Campbell, D. Gibbon, and D.J. Hirst, eds, *Proceedings of the 7th International Conference on Speech Prosody*, Trinity College, Dublin, Ireland, May 2014.