Journal of Speech Sciences 1(1):55-83. 2011.
Available at: http://www.journalofspeechsciences.org

THE ANALYSIS BY SYNTHESIS OF SPEECH MELODY: FROM DATA TO MODELS.

HIRST, Daniel [*]

CNRS & Université de Provence, Aix-en-Provence

This paper describes the application of the analysis by synthesis paradigm to the melody of speech. A complete chain of processes is described from the acoustic analysis of fundamental frequency (f_0), via the phonetic modelling of f_0 using the Momel algorithm, to the surface phonological representation of the curves using the INTSINT alphabet. Each step of the chain is designed as a reversible process which can be used to generate an acoustic output allowing an objective evaluation of the analysis. Finally, the current implementation of ProZed, a prosody editor for linguists, is described. It is argued that an explicit set of modelling tools like this will allow linguists to test different models of phonological structure which, it is hoped, will result in the availability of more and better data on a wide variety of languages.

Keywords Speech prosody; melody; intonation; analysis by synthesis

[*] Author for correspondance: daniel.hirst@lpl-aix.fr

1 Introduction

As a native speaker of English who has been living and working in France for the last 40 years, I often wonder what I could answer, if asked to summarise what I actually know today about the prosodic differences between the French and English languages, two languages with which I obviously know rather well. I could certainly give a talk or write an essay on the subject, but how many of the differences which I might describe could be called established facts?

And of course if the question asked concerned two languages with which I am much less familiar, like Korean and Finnish, for example, my answer would be even more reserved, since these are two languages which I do not speak or understand although they are, as it happens, languages in which I have been interested for several years and which I have often heard spoken.

So what, then, could I possibly hope to say about the prosody of the other 6000 odd languages that are said to exist in the world, the vast majority of which I have never even once heard spoken?

This seemingly impossible task is precisely what faces us when we try to answer questions like: *What is intonation? What is rhythm?* particularly so if the question is not what can we say about these topics, but what do we really know?

Having worked in the field of speech prosody for nearly as long as I have lived in France, I find it a sobering thought that our actual objective knowledge of the subject is still rather limited. We probably know quite a lot about segmental phonology today, but despite abundant publications and numerous theories, it seems that we still *know* rather little about speech prosody.

1.1 Linguists as human scientists...

Knowledge is the concern of Science, as the Latin for knowledge (*scientia*) suggests. Knowledge about language is the concern of linguistics, which often described as one of the human sciences.

There are many definitions of Science. One of the most attractive ones I know was given by the French physicist and Nobel laureate Jean-Baptiste Perrin (1870-1942) (Figure 1(a)), who defined the scientific method as:

> explaining visible complexity by invisible simplicity.
> (expliquer le visible compliqué par l'invisible simple.)

Scientists are confronted with the task of describing huge quantities of observable data. If they can reduce the complexity of the description by showing that the observable complexity is determined by some simpler, more abstract principle, then they have added to our knowledge of the data. Perrin's definition is in fact similar in spirit to the famous:

> Entia non sunt multiplicanda praeter necessitatem.
> (entities not are to-be-multiplied beyond necessity)

known as *Occam's Razor* and attributed to William of Ockham (1285-1349) (Figure 1(b)) and which we could rather freely paraphrase as:

> Don't use unnecessary variables in your description

More recently, the same fundamental idea has been given a formal basis under the name *Minimum Description Length* by the Finnish scientist Jorma Rissanen (Figure 1(c)), whose basic principle is that the best hypothesis for a given set of data is the one that leads to the best compression of the data (Rissanen, 1978).

(a) Jean-Baptiste Perrin (b) William of Ockham (c) Jorma Rissanen

Figure 1: Authors of three related definitions of science.

A good example of such a scientific method is given by the Russian chemist Dmitri Mendeleev (1834-1907), who, in 1869 developed a version of the tabular presentation of the Periodic Table of the Elements, to illustrate recurring (*periodic*) trends in the properties of the then-known elements. Mendeleev showed that what might, at the time, have appeared to be an arbitrary list of elements was in fact characterised by an underlying (invisible) structure, and that this structure accounted for many of the properties of the known elements in a systematic way. This paved the way to a better understanding of the atomic structure of the elements and not only predicted how many more elements were yet be discovered but also predicted what the properties of these elements would be.

Mendeleev's discovery illustrates nicely the properties of science, which is supposed to be *cumulative*, *explicit*, *predictive*, and *empirically testable*.

These are precisely the criteria which I believe we need to apply to linguistic knowledge.

One of the biggest criticisms which is often made of linguistic research is the fact that linguistic knowledge very often tends not be *cumulative* - each new theoretical framework, instead of building on previously established facts, defines a new set of criteria for what is a relevant theory.

One of the reasons for this is that, all too often, theories are not entirely *explicit*. Thus, most theories and descriptions of prosodic phonology, for example, are not developped to the point where they actually characterise measurable physical properties of speech acts.

A model which does characterise such properties will be *predictive*, since it will be possible to generalise from the observed data and make predictions about data which has not yet been analysed.

An explicit predictive model, then, will also be *empirically testable* since it will be possible to measure in some way the distance between the observed data and that predicted by the model.

The development and testing of explicit models of this type is the only way in which linguistics can hope to begin to accumulate knowledge rather than to speculate on the nature and relative elegance of abstract models. See also Xu (2011, this volume) for a detailed argument on the necessity for predictive models in the area of speech prosody. I argue that the best way to create an empirically testable model of speech prosody is through a process of analysis by synthesis.

The analysis by synthesis paradigm is potentially an attractive one for linguists, since it provides an empirical solution to the problem of validating an abstract model. The interaction between linguists and engineers has always been a productive area of exchange. This is particularly evident in the area of speech prosody. If the representation derived from a model can be used as input to a speech synthesis system, and if the contrasts represented in the model are correctly rendered in the synthetic speech, then the representation can be assumed to contain all the information necessary to express that contrast.

In a previous paper (Hirst et al., 2009), I described the application of this type of approach to the study of speech rhythm. In this paper, I describe the application of the approach to another important aspect of speech prosody: speech melody. The term *melody* is used here rather than the term *intonation* which is often used as a synonym for melody. As I have discussed in other publications, however (in particular cf. Hirst and Di Cristo, 1998b), I prefer to use the term intonation to refer to a more abstract system englobing both accentuation and rhythm as well as melody, while the term melody itself unambiguously refers to the modifications of pitch of an utterance over time. Much of what I have to say in this overview has already been published in various places but I try to give here an overall picture of the application of this approach and to provide answers to some of the questions which I have often been asked when I have given oral presentations of this material.

1.2 The analysis by synthesis paradigm

Analysis by synthesis involves trying to set up an explicit predictive model to account for the data which we wish to describe. A model, in this sense, is a system which can be used for analysis - that is deriving a (simple) abstract underlying representation from the (complicated) raw acoustic data. A model which can do this is explicit but it is not necessarily predictive and empirically testable. To meet these additional criteria, the model must also be *reversible*, that is it must be possible to use the model to synthesise acoustic data from the underlying representation. If we can do both these, as in Figure (2), then we can compare the output of the model to the original data and use this comparison to evaluate the model.

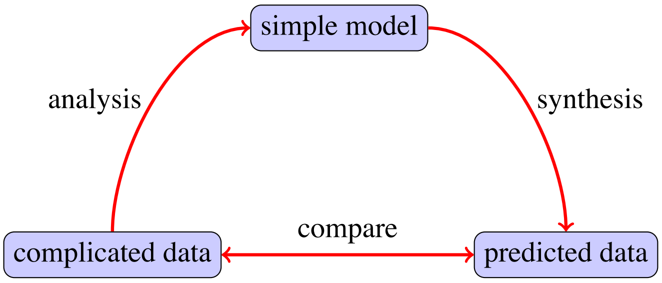


Figure 2: The Analysis by Synthesis paradigm

2 Modelling speech melody

Speech melody is essentially determined by the fundamental frequency (f_0) of the voiced parts of speech. Before we try to model this melody, we must ensure that the fundamental frequency which we are analysing is correctly detected.

2.1 Detecting f_0

Most speech analysis software, such as Praat, uses some default parameters defining the minimum and maximum values which are allowed for the f_0. In Praat (Boersma and Weenink, 2011), for example, these values, referred to as *Pitch Floor* and *Pitch Ceiling*, are set by default to 75 and 600 Hz respectively. Unfortunately these default parameters are rarely satisfactory. Figure 3 shows the first two sentences of recording A01 of the Aix-Marsec corpus (Auran et al., 2004) with fundamental frequency detected using the default parameters.

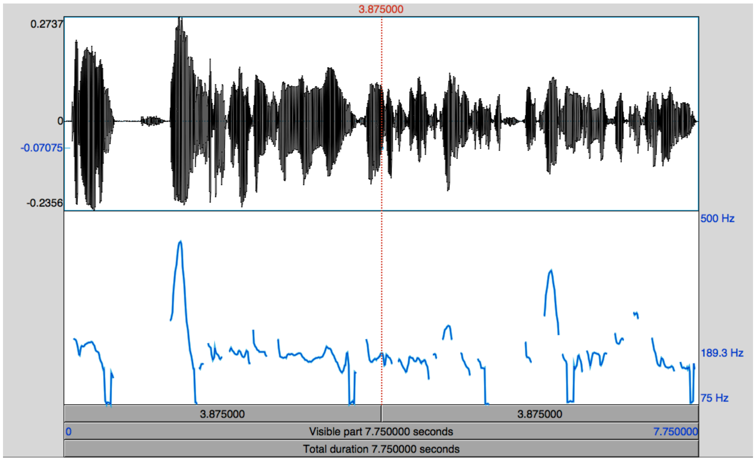


Figure 3: The first two sentences of recording A01 from the Aix-Marsec corpus. "Good morning. More news about the reverend Sun Myung Moon, founder of the Unification church, who's currently in jail for tax evasion." f_0 is detected using Praat's default parameters for F0. Range = [75:600 Hz]

The fundamental frequency of this extract contains some errors which are more apparent if we zoom in as in Figure 4 showing the end of the word 'morning'.

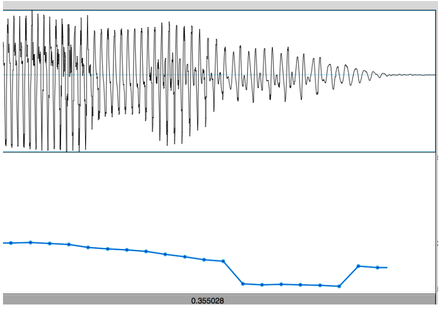


Figure 4: Detail of f_0 for the end of the word "morning."

The solution my students and I have adopted, and which is implemented in the Momel-Intsint plugin which I describe in more detail below, is a two-pass method. In the first pass, we use default parameters. The Pitch Floor is set at 50 Hz instead of Praat's default 75, since some male speakers go as low as 60 Hz but rarely much lower than that. The Pitch Ceiling is set at 700 Hz (instead of 600) which is generally high enough to include even very expressive female voices. Once we have calculated the fundamental frequency using these two parameters, we take the values of q_1 and q_3, the first and third quartiles of the f_0 distribution. It is well known that errors are likely to occur at the extreme values of a distribution but that the first and third quartiles of the distribution are much more robust estimates of the dispersion. Following empirical results established by De Looze (2010), we estimate new values for the Pitch Floor and Pitch Ceiling where the Pitch Floor is given by the formula $0.75 * q_1$ and the Pitch Ceiling is given by the formula $1.5 * q_3$.

Applied to the recording shown in Figure 3 this gives the values (rounded to the nearest 10 Hz) of 120 Hz and 300 Hz for the Pitch Floor and Pitch Ceiling respectively.

Unfortunately, as we can see from this example (Figure 5), this is not always sufficient to detect the f_0 correctly.

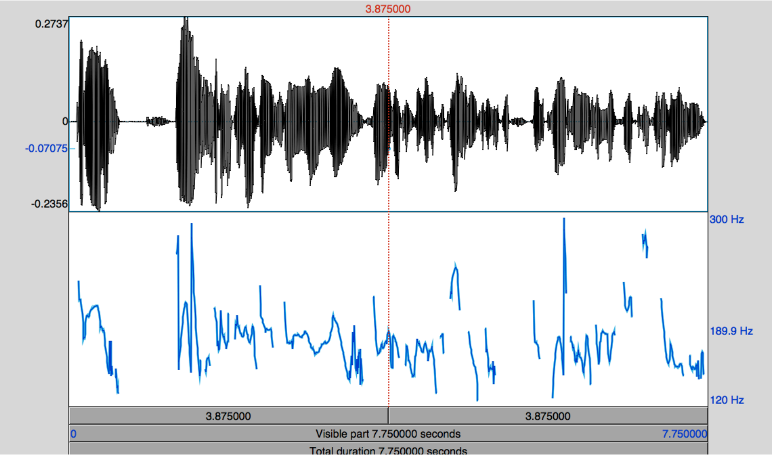


Figure 5: The same extract as in Figure 3. Parameters from the first implementation of the two-pass method. Range = [120-300]

The rise in pitch on the word "More" is so high in this fairly expressive speech style that it goes above the Pitch Ceiling we used, as can see even better if we zoom on the word "More" as in Figure 6.

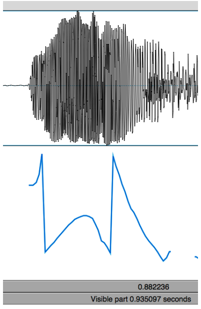


Figure 6: f_0 on the word "More". Example of octave error: pitch halving

In fact, there is a fundamental difference in the importance of the Pitch Floor and the Pitch Ceiling. Basically, we can say that if the Pitch Floor is too low then we are likely to get octave errors like that shown in Figure 4. Of course if the Pitch Floor is too high then we will get the opposite type of octave error, namely a doubling of the value of the measured f_0. For the Pitch Ceiling, if the value is too low, we will get errors like that in Figure 6. Setting the Pitch Ceiling too high does not, however, seem to lead to any systematic errors.

In the most recent implementation of the automatic pitch detection algorithm, then, the value of the Pitch Ceiling is set to a much higher value with the formula $2.5 * q_3$. This generally gives a good estimate of the f_0 of utterances, without the need for manual estimation of the floor and ceiling for the analysis, as can be seen in Figure 7.

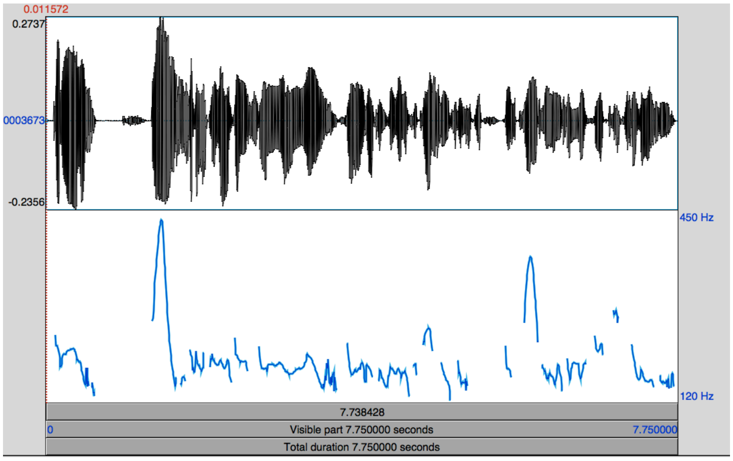


Figure 7: The same extract as in Figure 3. Parameters from the current implementation of the two-pass method. Range = [120-450]

2.2 Models of fundamental frequency

The search for an appropriate scale for measuring fundamental frequency has been one part of a systematic attempt, in particular by researchers from Holland (Hart , 't), 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 resynthesise the utterance, the result is judged to be perceptually equivalent to the original intonation pattern.

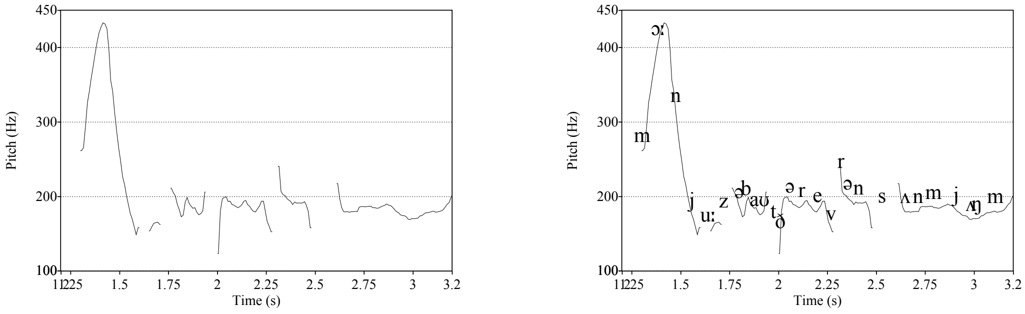
Another approach to modelling pitch has been to attempt to model the way in which pitch is produced by speakers. In particular, work by Fujisaki and his colleagues has applied a model of pitch production (Fujisaki, 1991) to a large number of languages, including several tone-languages, analysing an intonation 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 compatible with both. This is the approach which I follow in this presentation.

2.3 Micro-melodic effects

Fitting a raw f_0 curve with a mathematical model is not a simple straightforward problem due to the fact that fundamental frequency curves, as in the example we have just seen, are not always continuous: unvoiced portions of the utterance have no associated f_0. Even when the curve is continuous it is often not smooth and this type of irregularity is often hard to model simply.

If we look at a two second extract of the utterance which we have just seen, corresponding to the words "More news about the Reverend Sun Myung Moon" we notice (Figure 8(a)) that the beginning and end of the f_0 curve is in fact fairly continuous and smooth. The reason for this becomes obvious when we look at the phonemes associated with the curve as in Figure 8(b).



(a) raw f_0

(b) f_0 and phonemes

Figure 8: Two second extract of the f_0 curve corresponding to "More news about the reverend Sun Myung M(oon)".

The smooth portion at the beginning of the extract corresponds to the phonemes /mɔːnjuː/ whereas that at the end of the extract corresponds to /ʌnmjʌŋm/. All of these phonemes are fully sonorant, either vowels, semi-vowels or nasals. The discontinuity and irregularity of the f_0 curve is due to the presence of obstruents in the utterance: stops and constrictives, which either interrupt the curve (for voiceless obstruents) or make it irregular (for 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.

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 disturb or interrupt the flow.

The raw fundamental frequency curve then can be thought of as the interaction between two 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 corresponds to the observation (Nooteboom 1997) 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.

Linguists have known for a long time that fundamental frequency curves obtained from utterances containing only sonorants and vowels are much better behaved than raw f_0 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 (Gårding, 1998) or Annti Iivonen's "Laina lainaa Lainalla lainen" ('Laina lends Laina a loan') for Finnish (Iivonen, 1998). (cf Figure 9).

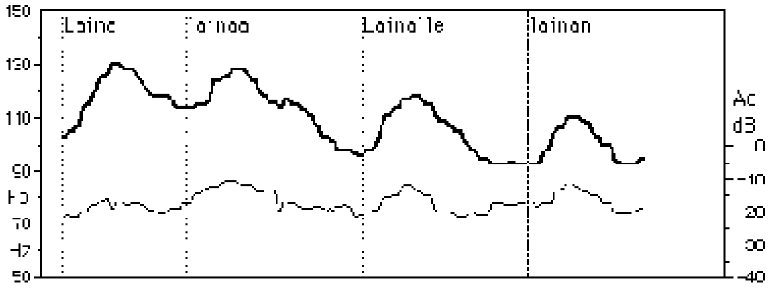
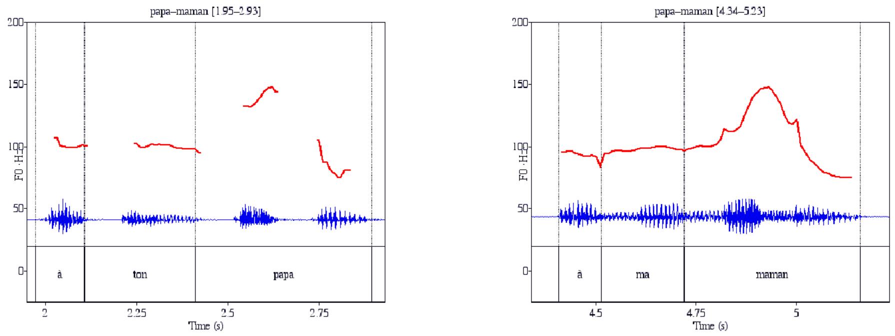


Figure 9: An example of a Finnish intonation pattern on a sentence with only sonorant phones. The sentence is "Laina lainaa Lainalla lainen" ('Laina lends Laina a loan'). (Iivonen, 1998)

2.4 Macromelody and micromelody

The idea, then, is that a raw intonation pattern is the interaction between two independent components: a macromelodic component determined by the accentuation and intonation of the utterance and a micromelodic component determined by the segmental phonemes. If we compare two simple utterances like "A ton papa." and "A ma maman." pronounced with a declarative intonation pattern, we can see that there is the same underlying macromelodic pattern for the two utterances and that the surface differences are simply due to the different phonemes of the utterances, voiceless stops in Figure 10(a) and sonorant nasals in Figure 10(b).



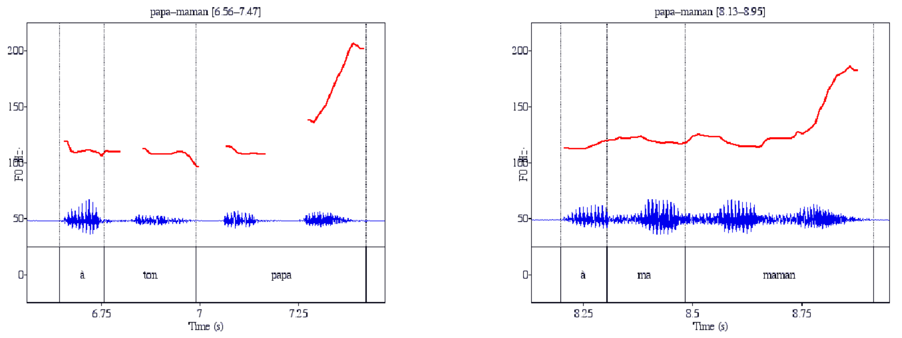
(a) A ton papa. ('to your Daddy')

(b) A ma maman. ('to my Mummy')

Figure 10: Two French sentences with a declarative intonation pattern

What is particularly worth noting is that the f_0 curve shown in Figure 10(a) is practically superposable on that of Figure 10(b). It seems as if the f_0 curve continues to change during the voiceless segments of the utterance even though it is not of course visible. This of course is not surprising if we think in terms of continuous changing of tension of the vocal folds which can of course continue to change even during voiceless segments.

If we now compare these patterns with those observed on the 'same' utterances pronounced with an interrogative intonation pattern as in Figure 11, we see once more that the two contours are practically superposable. And once again it looks as if the f_0 curve in Figure 11(a) continues to change during the voiceless segments of the utterance.



(a) A ton papa ? ('to your Daddy?')

(b) A ma maman ? ('to my Mummy?')

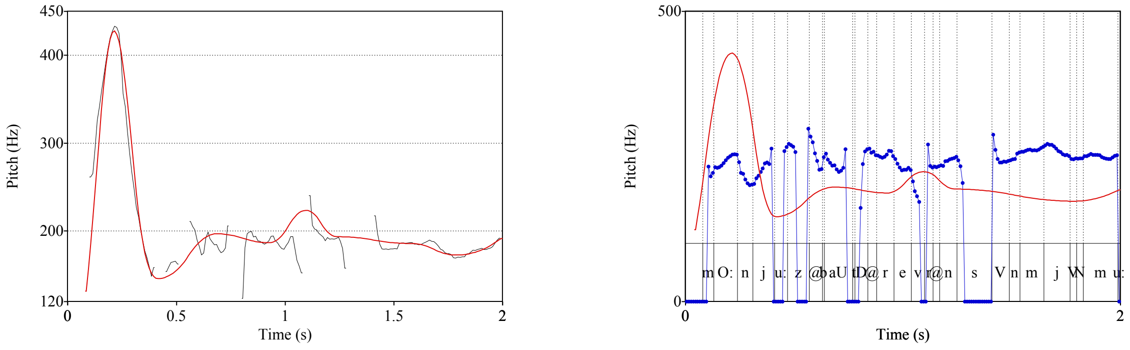
Figure 11: Two French sentences with an interrogative intonation pattern

Notice in particular that the final rise on 'papa' does not begin at the onset of the final vowel, the f_0 at this point is already considerably higher than that at the end of the preceding vowel. Notice that this idea of a continuously varying underlying pitch contour is not the model which is generally assumed in phonological descriptions of tonal and intonation contours. In the majority of these studies it is assumed that tones are directly associated with vowels (cf (Goldsmith, 1990) p. 44 for example) and that the fundamental frequency observed on the consonants is simply an interpolation between the tones on the vowels. If that were so, then it might be thought that the pitch curve visible in Figures 10(a) and 11(a) is actually closer to the underlying form than that in Figures 10(b) and 11(b) which are simply the result of an interpolation on the sonorant consonants. The fact that the f_0 curve follows the same trajectory in utterances with voiceless consonants as the smooth and continuous curve observed on the utterances with sonorants, however, and in particular the fact that the curve continues to evolve during the non-voiced portions of the utterance, seems to me convincing evidence that the planning of these curves is the result of an underlying macromelodic pattern on which the micromelodic variations are subsequently superposed.

2.5 A model for f_0 curves

2.5.1 Macromelodic and Micromelodic profiles

The macromelodic component of an intonation pattern, then, has, I shall assume, the two characteristics of being smooth and continuous. This is fortunate because, as I mentioned above, modelling a discontinuous or irregular function is much more difficult than one which is continuous and smooth. To return to the example we saw earlier, the underlying macromelodic profile of the curve might be something like the red curve in Figure 12(a) which might correspond to what we would produce if we were hum the sentence instead of pronouncing it. I return below to how the red curve was actually obtained; for the moment let us just assume that we have somehow obtained this macromelodic profile.



(a)

(b)

Figure 12: Raw f_0 (black) together with macromelodic (red) and micromelodic (blue) profiles for the first two seconds of recording A01

Once we have a macromelodic profile, we can derive the micromelodic profile by dividing each value of the raw f_0 curve by the corresponding value of the the modelling function. This gives a result as displayed in Figure 12(b), where the red curve is the macromelodic profile and the blue curve shows the micromelodic profile derived as just described.

Notice that such a modelling technique is not simply a stylisation of the raw f_0 curve, since the raw curve has actually been factored into two orthogonal components without any loss of information. Multiplying each value of the red curve from Figure 12(b) by that of the blue curve in the same figure gives the original raw f_0 as in the black curve in Figure 12(a).

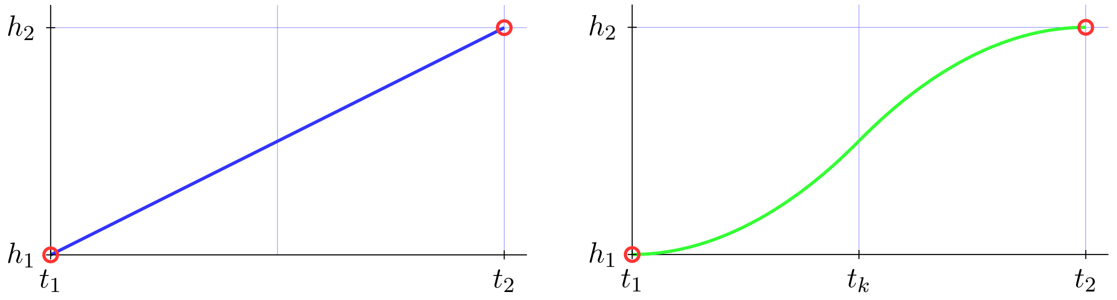
For speech synthesis it would of course be possible to model the micromelodic profile itself and to use this to improve the segmental quality of the utterance (for an application to Arabic see (Chentir et al., 2009)). For the study of intonation, the resynthesis of the utterance with the macromelodic profile is generally of sufficient high quality.

2.5.2 f_0 transitions

One of the simplest ways to model a smooth continuous function like that in Figure 12(a) is as piecewise sequence of transitions between successive points on the curve. I shall refer to these points, following a fairly long tradition, as *target points* even though it should be noted that this name is not intended to imply that the 'targets' necessarily have have any specific psychological reality for the speaker and listener. The advantage of a piecewise function over a global function is that each segment of the curve is defined locally by its own set of parameters, which means that a modification of one portion of the curve does not entail modifications throughout the rest of the curve. The simplest model, of course, would simply be a linear transition between two target points as in Figure 13(a) where the transition is defined by the function

$$h_i = h_1 + \frac{(t_2 - t_i)}{(t_2 - t_1)} \cdot (h_2 - h_1) \tag{1}$$

where h_1 and h_2 are the f_0 values of two adjacent targets points and where t_1 and t_2 are the corresponding time values of these targets.



(a) An linear f_0 transition

(b) A quadratic f_0 transition

Figure 13: Linear and quadratic transitions from a first target point $< t_1, h_1 >$ to a second $< t_2, h_2 >$
. Here the f_0 value (h_2) of the second target is higher than that (h_1) of the first target but the same reasoning would apply if it had been lower.

Naturally occurring f_0 curves, of course, are not linear but curvilinear. A number of mathematical functions have been used in the past to model such functions. One of the simplest of these functions is a quadratic transition, corresponding to a constant acceleration followed by a constant deceleration of the pitch change as shown in Figure 13(b) and as defined by the function:

$$\begin{aligned} t_i \in [t_1 \ldots t_k] : h_i &= h_1 + \frac{(h_2 - h_1) \cdot (t_i - t_1)^2}{(t_k - t_1)(t_2 - t_1)} \\ t_i \in [t_k \ldots t_2] : h_i &= h_2 + \frac{(h_1 - h_2) \cdot (t_i - t_2)^2}{(t_k - t_2)(t_1 - t_2)} \end{aligned} \tag{2}$$

As can be seen from Figure 13(b), in the case of a rise the transition consists of a concave curve from time t_1 to time t_k, the point of maximum slope, followed by a convex curve from t_k to t_2.

Figure 14 shows the same extract we have seen several times now, with the target points which define the curve represented as green circles.

2.6 Momel

A piecewise quadratic function such as that illustrated here is known as a quadratic *spline* function and has been in use in our laboratory since the 1980s to model intonation patterns using an algorithm called *Momel* (for "modelling melody").

The Momel algorithm is in fact formally equivalent to a subset of the contours which can be produced by the Rise/Fall/Connection (RFC) model of intonation later developed by Paul Taylor (Taylor, 1994) as a tool for speech synthesis. The only difference is that

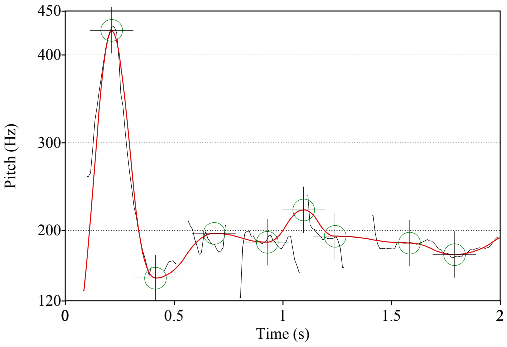


Figure 14: Macromelodic profile (red) for a two-second extract from recording A01, defined as quadratic transitions between target points (green).

the RFC model allows linear interpolations between two successive target points as well as quadratic interpolations. Of course if two successive target points have the same value of f_0 then the transition will be linear (i.e. flat) with Momel too. I have personally never observed a case where a non-flat linear transition gives a better approximation to an f_0 curve than a quadratic one.

The original implementation of Momel allowed the user to define target points manually by clicking on a representation of the f_0 curve on the computer screen. The user could then resynthesise the utterance using PSOLA resynthesis. This can be done today with Praat by creating a Manipulation object and removing and adding Pitch points manually. Praat displays the Pitch curve with linear interpolation between the Pitch points but an approximation of the quadratic spline function can be obtained by the command **Interpolate quadratically...**.

Manual modelling of f_0 is, of course, highly subjective and it was for this reason that my colleague Robert Espesser and I developped an automatic version of the algorithm (Hirst and Espesser, 1993), based on our experience of using the manual implementation of the model over a period of several years. The algorithm, which is described in detail in Hirst et al. (2000) uses a form of robust regression to optimise the modeling of raw fundamental frequency curves with a quadratic spline function.

The algorithm was later evaluated on a corpus of read speech in 5 languages (corpus Eurom1) during the course of the Multext European project (Véronis et al., 1994). Evaluators were instructed to correct the target points for the modelled speech only when such corrections made an audible improvement to the resynthesis of the speech. Table 1 shows the statistics of *recall*, *precision* and *F-measure* [1] for these corrections for the corpus of read speech together with those for a corpus of spontaneous speech in French.

[1] These are statistics commonly used in the field of information retrieval. In our context, *precision* is the percentage of all automatically detected targets which were considered correct, while *recall* is the percentage of all "correct" targets which were automatically detected. The *F-measure* combines these two values by taking the harmonic mean so that

$$F = 2 * \frac{precision * recall}{precision + recall}$$

. For more details see http://en.wikipedia.org/wiki/Precision_and_recall

Table 1: Results for the evaluation of the automatic Momel algorithm on read speech for 5 languages (English, French, German, Italian and Spanish and for spontaneous speech in French (corpus *Fref*). Columns show Total number of targets detected, number of targets added manually, number of targets deleted and the statistical measures of recall, precision and F-measure (see text). Data from Campione (2001)

Corpus	Lang.	No. of points			Evaluation		
		detected	*added*	*deleted*	*recall*	*precision*	*F-measure*
Eurom	en	8380	623	125	93.0	98.5	95.7
	fr	6547	423	130	93.8	98.0	95.9
	ge	13595	1145	506	92.0	96.3	94.1
	it	9475	337	330	96.4	96.5	96.5
	sp	8985	651	16	93.2	99.8	96.4
Fref	fr	9835	532	744	94.5	92.4	93.4

The results of the evaluation were very encouraging. The F-measures for the different languages showed a global efficiency of around 95% and even the corpus of spontaneous French showed an F-measure of 93.4% even though the algorithm had not at all been optimised for spontaneous speech.

Examination of the errors in the targets showed that one type of error in particular occurred systematically. This concerned a pitch rise before a silent pause where, frequently, the algorithm missed the final pitch of the rise entirely. An example of this type of error can be seen in Figure 15. This error is understandable since the algorithm uses a local modelling technique, fitting a parabola to portions of the curve. As can be seen in the example, the raw f_0 exhibits a concave portion corresponding to the acceleration of the pitch rise but hardly any convex portion corresponding to the decelerating portion of the rise. This is, in fact, the reason why the original algorithm fails to produce a final target.

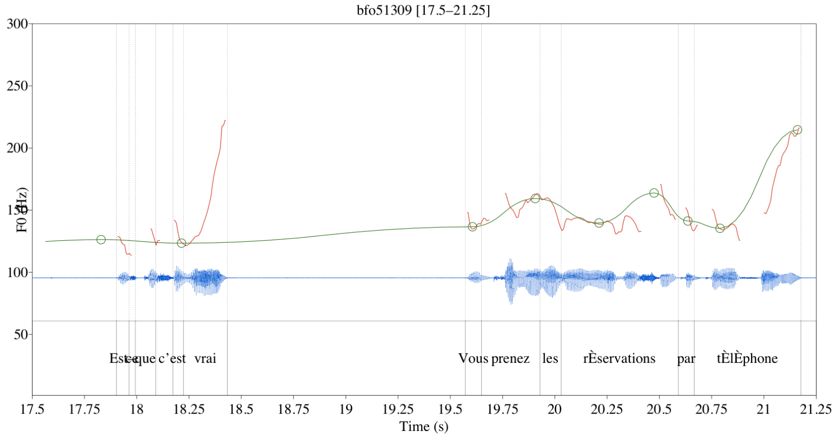


Figure 15: Old version of the automatic Momel algorithm for the utterance "Est-ce que c'est vrai ? Vous prenez les réservations par téléphone ?". Raw (red) and modeled f_0 (green).

The Momel algorithm has since been implemented as a Praat plugin (see Hirst (2007)) which allows users to use its functions directly from the Praat menus without needing to handle scripts directly. The systematic error we have just seen has been corrected by a special treatment before silent pauses. The concave part of the pitch rise is now extended to include a similar shaped convex portion. In other words, in order to obtain the best fit for this pitch rise, a high target point is calculated, situated in the silent pause and as near as possible to the previous low target. The target point is calculated so that the concave portion of the pitch rise follows the raw data points as closely as possible. The result of this improved algorithm can be see in Figure 16.

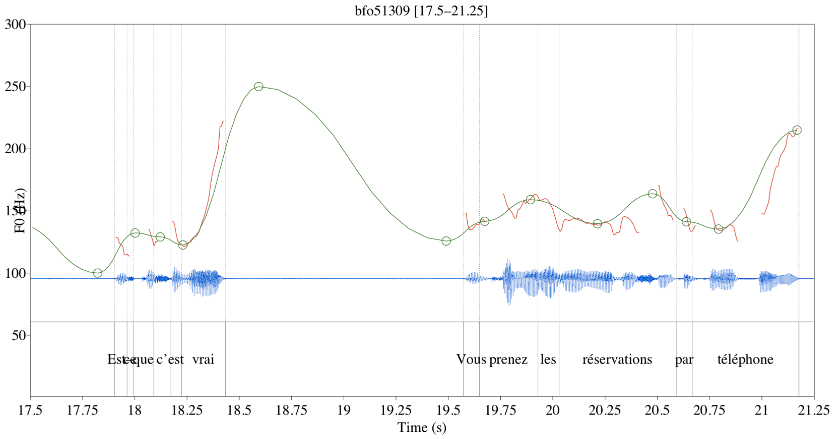


Figure 16: New version of the automatic Momel algorithm for the utterance "Est-ce que c'est vrai ? Vous prenez les réservations par téléphone ?" Raw (red) and modeled f_0 (green).

An evaluation of the improved algorithm was carried out on a corpus of read speech in Korean (Hirst et al., 2007). It showed a significant and systematic improvement as compared to the older version of the algorithm.

It is, naturally, desirable that the modelling tools we use should be as theory-neutral as possible. Complete neutrality, though, is obviously not entirely feasible, since any model necessarily makes some assumptions about the nature of underlying representations, as we saw above in the discussion of whether the underlying contour should be based only on the contours observed on the vowels or whether it should be modelled as a continuous underlying contour.

Rather than suggest that Momel is theory-neutral, then, I like to think that it is what we could call *theory-friendly*. I believe, that is, that the algorithm can be compatible with a number of different theoretical approaches to the description of speech melody. It has, in fact, been used in the past as a first step for modelling with the Fujisaki model (Mixdorff, 1999). It has also been used as first step for ToBI for both English (Maghbouleh, 1998; Wightman and Campbell, 1995) and Korean (K-ToBI) (Cho and Rauzy, 2008)).

It is also, of course, compatible with our own surface phonological representation alphabet, INTSINT which I describe below.

3 Coding Melody with INTSINT

3.1 INTSINT: an International Transcription System for INTonation

Unlike Momel, which provides a reversible modeling of the raw f_0 with no loss of information, INTSINT, an INternational Transcription System for INTonation was originally developed as a tool for linguists to provide a surface phonological representation of an intonation pattern. The original version of the system (Hirst, 1987) was based on an inventory of minimal pitch contrasts found in published descriptions of intonation patterns. The aim was to provide a tool for the systematic description of these intonation patterns, something along the lines of a narrow transcription using the International Phonetic Alphabet (IPA). Like the IPA, it was intended that INTSINT could be used for preliminary descriptions of intonation patterns, even for languages which had not previously been described. Notice that this aim is very different to that of the ToBI system (Silverman et al., 1992), which pre-supposes that the inventory of intonation patterns for the language being described has already been established. The official website for ToBI makes this particularly explicit:

> Note: ToBI is not an International Phonetic Alphabet for prosody. Because intonation and prosodic organization differ from language to language, and often from dialect to dialect within a language, there are many different ToBI systems, each one specific to a language variety and the community of researchers working on that language variety.
> (source: http://www.ling.ohio-state.edu/t̃obi/)

The INTSINT system (whose name was suggested to us by Hans 't Hart in a personal communication) was presented in Hirst and Di Cristo (1998a) where it was used for the description of 9 different languages. Basically, it describes an intonation contour as a sequence of Tonal segments which are labelled using an alphabet of 8 symbols. The tonal segments are assumed to be of three types:

Absolute tones *t*(op) *m*(id)*b*(ottom): These are assumed to refer to the corresponding position of the speaker's current pitch range.

Relative tones *h*(igher) *s*(ame) *l*(ower) Unlike absolute tones, relative tones are assumed to be refined with respect to the preceding tonal segment.

Iterative relative tones *u*(pstepped) *d*(ownstepped) These are also defined relative to the preceding tonal segment but generally involve smaller pitch changes and often occur in a sequence of steps either upwards or downwards.

In the chapters in Hirst and Di Cristo (1998a), the INTSINT tones were represented by graphic symbols represented between two horizontal lines and aligned with the text. These symbols were: Top [⇑]; Bottom [⇓]; Higher [↑]; Same [→]; Lower [↓]; Upstepped [<] and Downstepped [>]. The Mid tone was reserved for the unmarked onset of an Intonation Unit and was not marked. In most later publications the capital letters T, M, B, H, S, L, U, D were used instead of the graphic symbols. In this paper, the INTSINT tones are represented with lower case letters rather than upper case. This may help to avoid confusion with other more abstract coding schemes such as ToBI (Silverman et al., 1992), or the even more abstract underlying representation used in Hirst (1998), both of which use some of the same symbols as INTSINT.

3.2 Mapping from INTSINT to Momel

Although INTSINT was introduced as a descriptive tool for linguists, as its introduction was later than the creation of our f_0 modelling tool, we already had in mind the possibility that this *surface phonological* annotation could be linked to the analysis of the f_0 curve as a sequence of *phonetic* target points. This distinction between a *phonetic* representation and a *surface phonological* representation follows the distinction made by Trubetzkoy (1949) between *phonetics* which represents events using continuous variables and *phonology* which uses discrete categories. In this usage, it should be noted that what is generally called a 'phonetic representation' or a 'phonetic alphabet' is, more strictly, a surface phonological representation or alphabet, since the elements used for the representation are taken from a discrete set of symbols.

We anticipated, then, that it might be able to map the output of the *Momel* algorithm onto a sequence of symbols from the *INSINT* alphabet. In order to do this, following the idea of the analysis by synthesis paradigm as described above, it was first necessary to define a mapping in the other direction, that of synthesis. Some of the history of the way in which we defined this mapping is described in Hirst (2005). In its current implementation, the mapping depends on two speaker/utterance-specific parameters called *key* and *span* which together define the speaker's pitch range.[2]

The key (like a musical key) defines a central reference point for the speakers pitch range and the span defines the maximum and minimum pitch values of the range which are taken to be symmetrical (on a logarithmic scale) above and below the speaker's key.

The two parameters together define the three absolute tones, Top, Mid and Bottom, with respect to the speaker's pitch range as in the following formulas which assumes that the value of key is given in Herz and the value of span in octaves:

$$\begin{aligned} t &= key \times \sqrt{2^{span}} \\ m &= key \\ b &= \frac{key}{\sqrt{2^{span}}} \end{aligned} \tag{3}$$

The pitch targets corresponding to the relative tones are then defined with respect to both the preceding target (here called p) and the top (t) or bottom (b) of the range.

[2] note in some earlier publications I have used the word *range* for what I here refer to as *span*. I prefer now to use 'span' to refer to the interval, independently from the value of the 'key'. The two values 'key' and 'span' thus define the speaker's 'range'. So we might say that a given speaker has a pitch range from 100 Hz to 200 Hz, corresponding to a span of one octave and a key of 141 Hz.

A target point coded h is simply defined as the geometric mean (i.e. the mean on a log scale) of the preceding target and the top of the range - it thus corresponds to a pitch movement which moves up halfway towards the value of t. As can be expected, a target point coded s is defined as the same as the preceding target. Symmetrically to h, a target point coded l is defined as the (geometric) mean of the preceding target and the bottom of the range.

$$\begin{aligned} h &= \sqrt{p * t} \\ s &= p \\ l &= \sqrt{p * b} \end{aligned} \tag{4}$$

For the iterative tonal segments u and d the implementation defines the values as the (geometric) mean of the value of the preceding target and that which would be obtained if the target were coded h or l respectively. In other words these targets correspond to a pitch excursion one quarter of the way to the top/bottom of the pitch range.

$$\begin{aligned} u &= \sqrt{p * \sqrt{p * t}} \\ l &= \sqrt{p * \sqrt{p * b}} \end{aligned} \tag{5}$$

Assuming, once again, a logarithmic scale for the pitch range, these values are illustrated graphically in Figure 17.

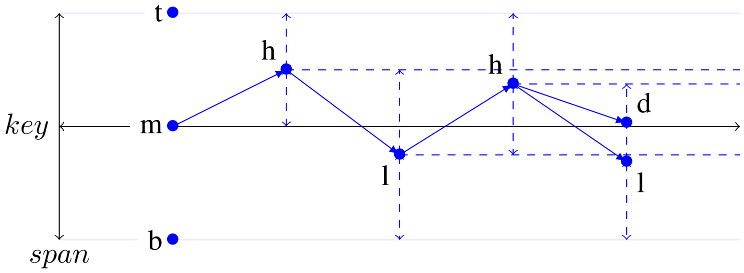


Figure 17: Graphic illustration of the mapping from INTSINT to Momel defined by 2 parameters key and $span$

This implementation obviously makes a number of assumptions, most of which would be open to empirical investigation. This of course, as I suggested above, is one of the major advantages of an explicit model such as this.

One consequence of the model, which was not specifically intended but which turns out to be fortunate, is that a sequence of alternating h and l tones will automatically introduce an iterative lowering of the tones, much like the probably universal effect of *downdrift* that has been described in the literature on tone and intonation as occurring in languages throughout the world. This downdrift effect, as illustrated in Figure 18, is thus an automatic by-product of the way in which the relative tones are defined.

Since it was first developed, the Momel algorithm has been applied relatively successfully to a number of different languages, including English, French, Italian, Catalan,

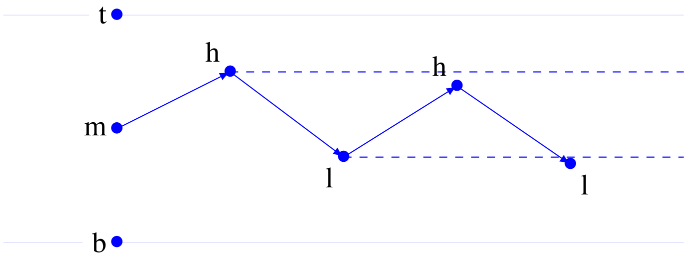


Figure 18: A graphic illustration of the fact that downdrift is an automatic by-product of the way in which the relative tones are defined

Brazilian Portuguese, Venezuelan Spanish, Russian, Arabic, isiZulu and Korean (for references see (Hirst, 2007)). More recently (Zhi et al., 2010), 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. An attempt to optimise window-size for the algorithm showed no overall improvement with respect to the manually corrected data. This was taken to confirm the fact (as had been suggested by Xu and Sun (2002)) that pitch change by speakers of a lexical tone language like Chinese is not notably faster than that produced by speakers of languages with no lexical tone. The annotated data obtained during this application will 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.

3.3 Mapping from Momel to INTSINT

Having defined a mapping from tonal segments (INTSINT) to pitch targets (Momel), the same model can be used to establish a reverse mapping from the targets to the tonal segments. As is generally the case, such a reverse mapping from continuous variables to discrete categories is much less straightforward than the mapping from categories to continuous variables. The approach we have adopted is an exhaustive search of the target space for the optimal values of the two parameters key and $span$ together with the optimal coding of the sequence of target points, given those two parameters.

The procedure as described in Hirst (2005) has been implemented as a Perl script. It assumes that the relevant target space is defined as follows:

$$\begin{aligned} key &= mean \pm 50 Hz \quad (step : 1) \\ span &= 0.5 \ldots 2.5 octaves \quad (step : 0.1) \end{aligned} \tag{6}$$

The script thus tries each of the possible values of the two parameters within this target space. For each of the 2000 possible couples $\langle key, span \rangle$ the script evaluates every possible coding of the target points using the formulas in equations (3, 4, 5) and calculating the sum of the square of differences between the predicted value and the observed value. The output of the script is thus the optimised value (within the target range) of the parameters *key* and *span* together with the optimal INTSINT coding using these parameters. The output of the script is a text file such as the following corresponding to the application of the script to the same extract which we saw earlier:

```
; A01_01.intsint created on Tue Aug 24 08:12:47 2010 by intsint.pl 2.11
; from A01_01.momel
;    32 values  mean = 191
<parameter span=1.4>
<parameter key=235>
0.113 M 221 235
0.219 D 205 208
0.434 D 182 190
0.746 B 120 145
1.177 S 120 145
1.423 T 428 382
1.623 B 146 145
1.894 U 197 184
```

…

Figure 19: Extract of a sample output of the automatic INTSINT coding script. After the values for *key* and *span* each line gives the time value (in seconds), the optimised INTSINT code, the original target point used as input and the predicted target point derived from the coding with the current values of *key* and *span*

In this output, the optimised values of key and span (here 235 Hz and 1.4 octaves respectively) are given, together with the sequence of target points. Each line gives the time value (in seconds) for the target, the optimised INTSINT code, the original target point used as input and the predicted target point derived from the coding with the current values of *key* and *span*

Unlike with Momel, there is some loss of information with the INTSINT coding (as can be seen from the differences between the 3rd and 4th columns of the output in Figure 19 the model is still however a reversible one which can be used for synthesis. Figure 20 shows the result of resynthesising the pitch targets from the results of the automatic INTSINT analysis applied to a 5 sentence passage from the Eurom1 corpus, compared to the targets obtained from the application of the Momel algorithm.

3.4 Longer term characteristics of pitch range

The implementation of INTSINT as described above presupposes that there are no variations in *key* and *span* within the segment of speech which is analysed. In authentic speech, such changes naturally occur quite frequently and are obviously very significant and important. One solution is to implement the INTSINT coding on smaller segments of speech such as breath groups, making what seems a fairly reasonable assumption that changes of *key* and/or *span* are more likely to occur *between* breath groups rather than

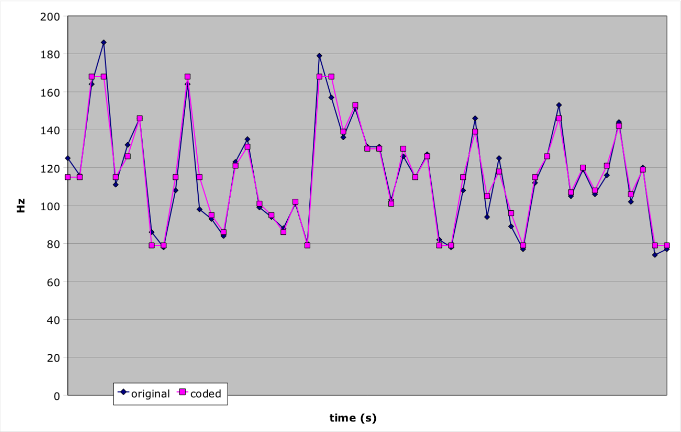


Figure 20: Predicted (pink) versus observed (blue) values for pitch targets for a 5 sentence passage after automatic coding with the INTSINT alphabet

within them. For an investigation of the possibility of automatising such a process, cf. (De Looze, 2010) which implements an algorithm (AdoReVa) applying a cluster analysis for this task (see Figure 21).

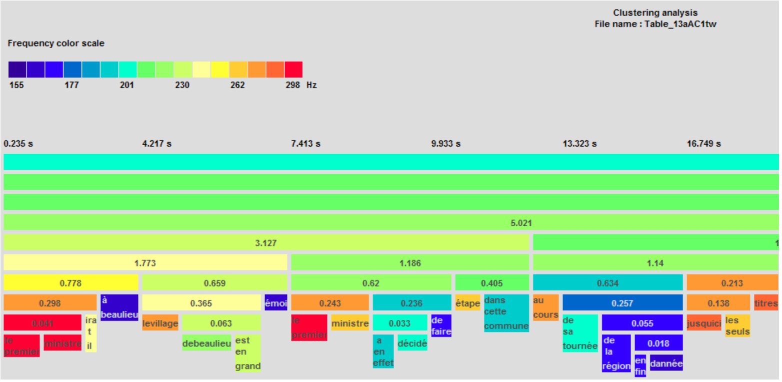


Figure 21: The AdoReVA algorithm (De Looze, 2010), a cluster analysis for the automatic detection of changes of pitch register.

In this analysis, instead of providing a sequence of boundaries, the algorithm provides a hierarchy of potential boundaries organised as a layered icicle diagram as in Figure 21. The higher boundaries on the diagram indicate stronger boundaries. A colour scale indicates the key for each unit. The darker the colour, the higher the key.

4 ProZed

The availability of explicit models of speech melody such as those described above makes it possible to use such models to implement more abstract representations of prosody. With this aim in mind our current work in progress concerns the implementation of a *Prosody Editor for Linguists* which we have called *ProZed* and which aims to provide linguists with a tool with which they can experiment with different abstract phonological models, providing them with an acoustic output with which they can at least informally evaluate the relative value of different models.

A version of this editor applied to speech rhythm is described in Hirst and Auran (2005) which implemented an empirical linear model for rhythm where each *rhythm unit* is characterised by three parameters: q a (possibly speaker dependent) parameter defining a unit of quantal lengthening, t a long term parameter of *tempo*, and k a local scalar effect of lengthening specific to each rhythm unit. With these parameters the duration of the rhythm unit ρ is defined as:

$$\hat{d_\rho} = t \cdot \{\sum_{i=1}^{m} \bar{d}_{i/p} + k \cdot q\} \tag{7}$$

where $\bar{d_{i/p}}$ is the mean duration of all the phones labelled as the same phoneme p as that that occuring in position i in the rhythm unit.

ProZed will be implemented as a plugin to the Praat software (Boersma and Weenink, 2011). It will allow the manipulation of the rhythmic and the tonal aspects of speech as defined on three specific tiers in addition to the *phoneme* tier. The first two of these are named the *RU* (rhythm unit) tier and the *TU* (tonal unit) tier. These two tiers control the short term variability of prosody. Longer term variations will be controlled via a third tier named the *IU* (intonation unit) tier.

The speech input to the program may be natural recorded speech, the prosodic characteristics of which will then be modified by the software, or, alternatively it may be the output of a speech synthesis system with, for example, fixed durations for each speech segment.

The current version of the program is designed as the re-synthesis step of what is planned to be a complete analysis by synthesis cycle. This will be directly integrated with the output of the Momel-INTSINT and ProZed Rhythm analysis models which are described above.

In the following section I present the current state of the software for the modelling of speech melody.

4.1 ProZed melody

The annotation of the melody of an utterance is encoded via two interval tiers. These are the tonal tier *TU* and the intonation unit tier *IU*.

While it is hoped that linguists will find these tiers appropriate and useful levels for modelling the melody of speech, no assumptions are made as to the theoretical status of these units. Tonal units, are consequently *defined* as the domain of short term pitch variation and intonation units are *defined* as the domain of longer-term variation. For different linguists, these units may correspond to different phonological entities - the software provides a means to implement different interpretations in order to evaluate the effect of the different choice of units.

4.2 Determining pitch via the Tonal Unit (*TU*) tier.

Pitch in ProZed is determined by a representation of the contour using the *INTSINT* alphabet described above in section 3. The pitch height of a target is determined by the symbolic "tonal" symbol from the INTSINT alphabet which, as we saw, is defined either globally with respect to the speaker's current register or locally, with respect to the preceding target.

The actual fundamental frequency of the pitch targets is determined by the formulas 3, 4 and 5.

The timing of the target points is assumed to be determined with respect to the boundaries of the corresponding TU. In previous work (e.g. (Hirst, 1999)), I suggested that such timing might be limited to a restricted inventory of positions within the TU, such as initial, early, mid, late and final.

In this implementation, I adopt a more general solution and allow, in fact, an arbitrary precision of alignment via the use of "dummy" targets represented by the symbol "-". Using this annotation, a tonal target X which occurs alone in the middle of a unit will be coded /X/. When there are more than one tonal target in a *TU* then they are assumed to be spread out evenly, so that /W X/ will have one target occurring at the first quarter of the duration and one at the third quarter of the duration. This has the result that for two consecutive TUs each containing two targets, each successive pair of targets will be equally spaced apart. In order to represent a target at the third quarter of the duration with no preceding target the annotation /- X/ can be used. The symbol "-" is thus used to influence the timing of the other target but does not itself correspond to a pitch target.

The formula for calculating the timing of the ith target of a sequence of n targets in a TU beginning at time $start$ and ending at time end is:

$$t_i = start + \frac{(2i-1)}{2n} * [end - start] \tag{8}$$

In practice, I assume that a linguist will probably make a fairly sparse use of these dummy symbols but the annotation in fact allows the specific timing of a target or targets to be coded to an arbitrary degree of precision. Thus a representation like /- - X - - Y - Z - - - - -/, for example, could be used to specify timing very precisely, where in this case the targets would occur at 0.192, 0.426 and 0.577 of the duration of the interval, respectively (by applying equation 8, e.g $0.192 = (2*3-1)/(2*13)$). The actual precision of the timing is consequently left to the user to determine. It is particularly interesting to use an annotation system which can be rendered as precise or as general as wished so that the same annotation can be used in the analysis and in the synthesis steps of the analysis-by-synthesis procedure. In line with the concept of Minimal Description Length mentioned in section 1.1, the annotation gives an advantage to simpler representations over more detailed and complex ones.

4.3 Defining long term parameters with the Intonation Unit (*IU*) tier

The short term values obtained from the RU and TU tiers are finally mediated by the long-term parameters defined on the IU tier. These are currently *key* and *span* for pitch.

The two parameters are initialised with default values

key = 150 span = 1

where the value for key is specified in Herz and that for span is specified in octaves. Subsequently either of the values can be modified for the following IUs by simply including a specification of the value of the corresponding parameter or parameters, e.g.

span = 0.8

on the IU tier.

Each modification of a long-term value remains valid until it is modified at a later point in the representation.

The script also allows the definition of pitch targets at the extreme ends of an IU, which corresponds essentially to the target points interpreted as "boundary tones" in many phonological prosodic models.

This is done using an annotation such as /$[mb]$/, for example, which will place a /m/ target point located at the beginning of the unit and a /b/ target located at the end. It is also possible to specify only one boundary tone, so that /$b]$/ will place only a bottom target at the end of the unit with nothing at the beginning whereas /$[m$/ will place a target at the beginning of the unit with nothing at the end.

This implemenation corresponds, in fact, to a non-linear phonological model similar to that described in Hirst (1998), where tones are aligned with Tonal Units and Intonation Units and where syllables are also aligned with Tonal Units and Intonation Units, but where the tones and syllables are not ordered with respect to each other. Figure 22, for example, illustrates a possible surface phonological representation for the utterance "My friend had to go to the doctor's.", where τ and σ represent the tones and the syllables, respectively.

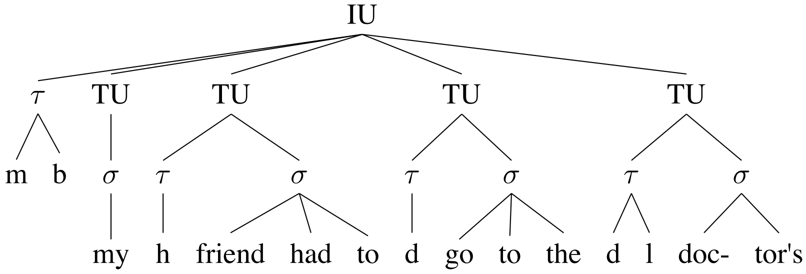


Figure 22: A possible surface phonological representation of the utterance "Last week, my friend had to go to the doctor's." τ corresponds to tones and σ to syllables.

This could be encoded using the ProZed application as in Figure 23.

The TextGrid shown in this Figure is the direct implementation of the representation in Figure 22, together with the indication of the key (102 Hz) and $span$ (0.9 octaves) for the utterance. Running the script performs the task of linearising the tonal segments with respect to the IU and TUs (and hence of course with respect to the syllables and phones) and generates the corresponding target values for each tonal segment, adding

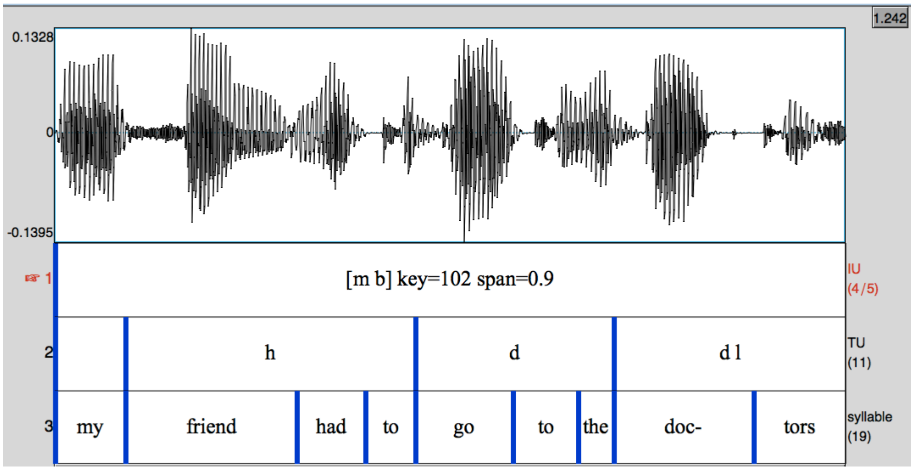


Figure 23: Using the ProZed script to model melody

the aligned tonal symbols and the corresponding target points as tiers in the TextGrid as in Figure 24.

The original utterance can then be re-synthesised with the pitch contour generated from the target points.

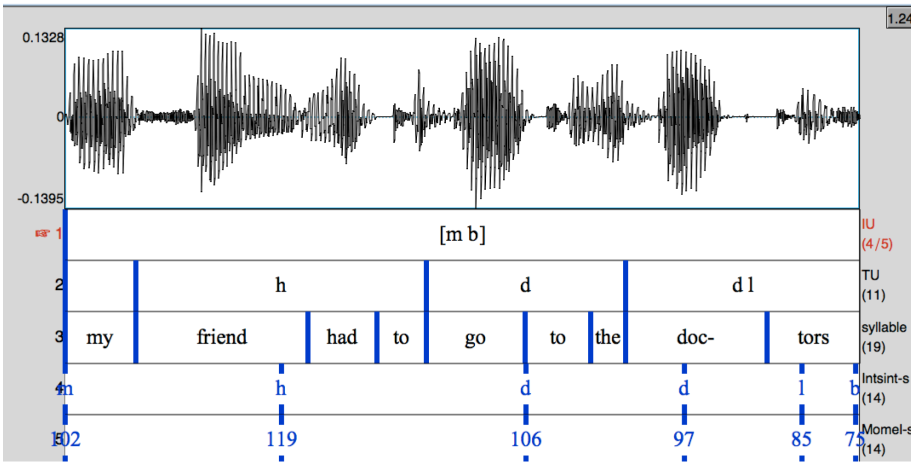


Figure 24: Output of the ProZed script applied to the TextGrid in Figure 23

4.4 Integrating the synthesis with the automatic analysis of pitch

The output of the Momel and INTSINT analysis algorithms, as described above, can be directly used as input to the re-synthesis module as described above although at present this has to be done manually since the tonal segments and target points in the current implementation are aligned directly with the acoustic signal rather than being aligned relative to the IUs and TUs. Nevertheless, this implementation makes it extremely simple for a user to create a TextGrid like that in Figure 23 and then to test different variants of the coding and the alignment, in order to see which aspects sound most significant.

The current implementation of the Praat script *ProZed-melody* is available at:

http://uk.groups.yahoo.com/group/praat-users/files/Daniel_Hirst/

Updates and extensions to the script will be made available at the same address, as, eventually, will the plugin version of the scripts used in the ProZed environment.

4.5 Using prosodic form to discover prosodic functions

As I have often stressed in previous work, (cf. in particular (Hirst, 2005)) I believe it is essential for a prosodic analysis to make a systematic distinction between prosodic *form* and prosodic *function*. In a recent study of Finnish (Vainio et al., 2009), it was shown that a very high quality synthesis of a continuous text in that language could be obtained without any annotation of prosodic form, by bootstrapping a functional annotation applied manually to about 20 sentences and then predicting the functional annotation on a larger corpus by multilinear regression from the acoustic data. The functional annotation used in this application consisted basically of four degrees of prominence (unaccented, accented, nuclear accent and emphatic accent) and three degrees of boundary (none, minor and major).

This result could be taken to mean that there is no need for an explicit model of prosodic form in the task of improving the quality of speech synthesis. I believe, however, that there is a crucial role for such a model in establishing and enriching the inventory of prosodic functions.

By developing a mapping from prosodic functions to prosodic forms, we can evaluate the accuracy of the predictions from the inventory of functions. Applying this to a corpus of speech would then allow us to measure the discrepancies between the predicted and the observed data, which in turn will allow us to extend the inventory of functions in order to reduce these discrepancies. For a recent study applying this approach to English, see (Ali and Hirst, 2009). In her doctoral thesis in progress (Zhi, in progress), the first author of (Zhi et al., 2010) has shown that a satisfactory re-synthesis of spontaneous utterances in Standard Beijing Chinese can be obtained from a surface phonological representation using the INTSINT alphabet and corresponding to a mapping from the representation of the lexical items with their associated lexical tones, even though , the details of this mapping, have yet to be firmly established.

5 Conclusion

This paper has described in some detail the application of the analysis by synthesis paradigm to the task of analysing one important component of the intonation of speech, speech melody. Of course, although I have concentrated on the description of one particular model, much of what I have said in this presentation would apply equally well to other recent explicit models of speech melody such as Rosenberg (2010) in the ToBI framework, or Prom-on et al. (2009) for a model based on physiological and functional criteria.

Although speech technology has become more and more accessible in recent years, it remains nonetheless true that the gap between application and users is still far too wide.

This is unfortunate since there are a great number of linguists throughout the world who are particular interested in developing and testing different models of phonological structure.

Providing linguists with better tools will surely result in the availability of more and better data on a wide variety of languages, and such data will necessarily be of considerable interest not only to linguistics as a potentially cumulative science but also to engineers working with speech technology. I sincerely hope that the ProZed algorithm which I describe here will make a modest contribution to this development.

REFERENCES

Ali S, Hirst D. Developing an automatic functional annotation system for british english intonation. In Proceedings of Interspeech X. Annual Conference of the International Speech Communication Association. Brighton, 2009.

Auran C, Bouzon C, Hirst D. The Aix-MARSEC Project: An Evolutive Database of Spoken British English. Speech Prosody 2004, International Conference, March 23-26 2004, Nara., 2004.

Boersma P, Weenink D. Praat: doing phonetics by computer [computer program]. 2011.

Campione E. Etiquetage semi-automatique de la prosodie dans les corpus oraux - algorithmes et méthodologies. Ph.D. thesis, Université de Provence, 2001.

Chentir A, Guerti M, Hirst D. Extraction of standard arabic micromelody. Journal of Computer Science, 5(2):86--89, 2009.

Cho H, Rauzy S. Phonetic pitch movements of accentual phrases in korean read speech. In Proceedings of the 4th International Conference on Speech Prosody. Campinas Brasil., 2008.

De Looze C. Analyse et interprétation de l'empan temporel des variations prosodiques en français et en Anglais. Ph.D. thesis, Université de Provence, Aix-en-Provence, France, 2010.

Fujisaki H. Modeling the generation process of F0 contours as manifestation of linguistic and paralinguistic information. In Proceedings of the XIIth International Congress of Phonetic Sciences, pages 1--10. 1991.

Gårding E. Intonation in swedish. In D Hirst, A Di Cristo (editors), Intonation Systems. A Survey of Twenty Languages., chapter 6, pages 117--136. Cambridge: Cambridge University Press, 1998.

Goldsmith JA. Autosegmental and metrical phonology. Cambridge, Mass.: B. Blackwell, 1990.

Hart ('t) J, Collier R, Cohen A. A perceptual study of intonation: an experimental-phonetic approach to speech melody. Cambridge University Press, 1990.

Hirst D. La représentation linguistique des systèmes prosodiques : une approche cognitive. Thèse de Doctorat d'Etat (Habilitation Thesis), Université de Provence, 1987.

Hirst D. Intonation in British English. In D Hirst, A Di Cristo (editors), Intonation Systems. A Survey of Twenty Languages., chapter 3, pages 56--77. Cambridge: Cambridge University Press, 1998.

Hirst D. The symbolic coding of segmental duration and tonal alignment: an extension to the intsint system. Sixth European Conference on Speech Communication and Technology, 1999.

Hirst D. Form and function in the representation of speech prosody. Speech Communication, 46(3-4):334--347, 2005.

Hirst D. 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, pages 1233--1236. Saarbrucken, 2007.

Hirst D, Auran C. Analysis by synthesis of speech prosody: the prozed environment. In Proceedings of Interspeech 2005. (Lisbon), pages 3225--3228. 2005.

Hirst D, Bouzon C, Auran C. Analysis by synthesis of British English speech rhythm: from data to models. In G Fant, F Hiroya, S Jiaxuan (editors), Frontiers in Phonetics and Speech Science. A Festschrift for Professor Wu Zongji's 100th Birthday., pages 251--262. Beijing, Peoples Republic of China: Commercial Press, 2009.

Hirst D, Cho H, Kim S, Yu H. Evaluating two versions of the momel pitch modeling algorithm on a corpus of read speech in korean. In Proceedings of Interspeech, volume VIII, pages 1649--1652. Antwerp, Belgium, 2007.

Hirst D, Di Cristo A. Intonation Systems: A Survey of Twenty Languages. Cambridge University Press, 487 p., 1998a.

Hirst D, Di Cristo A. A survey of intonation systems. In D Hirst, A Di Cristo (editors), Intonation Systems: A Survey of Twenty Languages, chapter 1, pages 1--44. Cambridge University Press, 1998b.

Hirst D, Di Cristo A, Espesser R. Levels of representation and levels of analysis for the description of intonation systems. In M Horne (editor), Prosody: Theory and Experiment. Studies Presented to Gösta Bruce., pages 51--87. Kluwer Academic Pub, 2000.

Hirst D, Espesser R. Automatic modelling of fundamental frequency using a quadratic spline function. Travaux de l'Institut de Phonétique d'Aix, 15:75--85, 1993. URL `http://www.isca-speech.org/archive/eurospeech_1989/e89_1480.html`.

Iivonen A. Intonation in Finnish. In D Hirst, A Di Cristo (editors), Intonation Systems. A Survey of Twenty Languages, chapter 17, pages 331--347. Cambridge University Press, 1998.

Maghbouleh A. Tobi accent type recognition. In Proceedings of ICSLP., Paper 0632. 1998.

Mixdorff HJ. A novel approach to the fully automated extraction of fujisaki model parameters. In Proceedings of ICASSP 1999. 1999.

Prom-on S, Xu Y, Thipakorn B. Modeling tone and intonation in mandarin and english as a process of target approximation. Journal of the Acoustical Society of America, 125(1):405--424, 2009.

Rissanen J. Modeling by shortest data description. Automatica, vol. 14:465--471, 1978.

Rosenberg A. AuToBI -- a tool for automatic ToBI annotation. In Proceedings of the International Conference on Spoken Language Processing. 2010.

Silverman K, Beckman M, Pitrelli J, Ostendorf M, Wightman C, Price P, Pierrehumbert J, Hirschberg J. TOBI: A Standard for Labeling English Prosody. In Second International Conference on Spoken Language Processing, pages 867--870. Banff. Canada.: ISCA, 1992.

Taylor P. The rise/fall/connection model of intonation. Speech Communication, 15(1-2):169--186, 1994.

Trubetzkoy. Grundzüge der Phonologie. (French translation by J. Cantineau 1957) Principes de phonologie. Paris: Klincksieck, 1949.

Vainio M, Hirst D, Suni A, De Looze C. Using functional annotation for high quality multilingual, multidialectal and multistyle speech synthesis. In Proceedings SPECOM, 13th International Conference on Speech and Computer. St Petersburg, Russia, 2009.

Véronis J, Hirst D, Ide N. NL and speech in the MULTEXT project. In Proceedings of AAAI Workshop on Integration of Natural Language and Speech, pages 72--78. Seattle, USA, 1994.

Wightman C, Campbell N. Improved labeling of prosodic structure. In IEEE Trancactions on Speech and Audio Processing. 1995.

Xu Y. Speech prosody: a methodological review. Journal of Speech Sciences, 1(1):85--115, 2011.

Xu Y, 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.

Zhi N. The music of Beijing Chinese speech. On the interactions of tones and intonations in read and spontaneous Beijing speech. Ph.D. thesis, Scuola Normale da Pisa, in progress.

Zhi N, Hirst D, Bertinetto PM. Automatic analysis of the intonation of a tone language. applying the momel algorithm to spontaneous standard chinese (beijing). In Proceedings of Interspeech XI. Makuhari, Japan, 2010.