A Numerical Model of Pitch Perception for Short-Duration Vocal Tones: Application to Intonation Analysis

Christophe d'Alessandro
LIMSI - CNRS
BP 133, F-91403 Orsay, France

ABSTRACT

An algorithm for automatic intonation analysis is described. It is based on a two-parameter model of weighted time-averaging with threshold for pitch perception. This model can be considered as a non-linear filter. In a first stage speech is decomposed into short-duration tonal segments using short-term energy. In a second stage these short-duration tones are analyzed using the numerical model. A set of short (static and dynamic) tones are then obtained, together with their (constant or time-varying) pitches. Stylized FØ contours are reconstructed from this set of tones. Stylized contours are resynthesized, and give synthetic sentences which are perceptually identical with natural sentences.

INTRODUCTION

This paper presents a numerical model of pitch perception for automatic intonation analysis. Automatic intonation analysis is an important challenge for both fundamental and applied prosodic studies. FØ, as it appears at the output of a pitch tracker, is generally difficult to interpret. A similar situation is encountered in singing: although the melody is precisely indicated in the vocal score, is precisely realized by the singer, and is precisely appreciated by the audience, there is no hope to read the melody directly on FØ curves. Perceptual constraints on melodic accuracy are more severe in singing than in speech, but in singing like in speech, many details of FØ are not perceived as significant variations, and many other details are not perceived at all. Clearly, a quantitative model of pitch perception is needed in order to fill the gap between the output of pitch trackers and the tonal decomposition described by linguistic analyses.

A quantitative model of tonal analysis for French has been presented by Mertens (1987). Both a linguistic description of French intonation and a system for automatic recognition of intonation have been developped. In the work by House (1990), qualitative results on the influence of spectral changes on intonation perception have been proposed, and a system for automatic recognition of intonation in Swedish has been described, along with an automatic FØ stylization procedure.

Compared to these works, we shall present herein a system using new psychoacoustic data on pitch perception for short tones. Our experimental data on pitch perception for short-duration tones with changing frequency were obtained in the context of a study on vibrato tones in singing. There is no room here to recall the experimental conditions of this study, and the reader is refered to d'Alessandro & Castellengo (1993) for details. We think that the model obtained for singing is fully applicable to intonation analysis in speech, because: 1) the durations, extents and FØ patterns used in our experiments are comparable to those observed in speech; 2) the psychological thresholds are probably higher for speech perception than for musical perception; 3) these thresholds are also higher for short-tones in isolation, compared to short-tones in context (see for instance Watson & al. (1990)).

TONAL SEGMENTATION

The main question in applying the results obtained for short tones to speech analysis is that, apparently, speech is not made of a succession of short tones. Nevertheless, it is generally accepted that a syllabic segmentation takes place in prosodic perception (at least for languages like French), and therefore one can consider that speech is perceived as a succession of tonal segments related to syllables.

An extensive review of syllabic segmentation, particularly in French, is reported in Mertens (1987): in this study the segmentation algorithms were mainly based on

short-term energy, or on loudness. Another criterion for syllabic segmentation has been proposed by House (1990): the so-called "spectral constraint hypothesis". Nevertheless, in the computer implementation described by House, it appeared that the segmentation component used intensity measurements in much the same way as Mertens. As far as we know, no quantitative psychoacoustic data are yet available to give some evidence on optimal threshold values. In our system, tonal decomposition is also based on ad-hoc energy thresholds. Provided that speech is decomposed into short-duration tonal segments, the model of short-tone perception can be introduced.

STATIC AND DYNAMIC TONES

In d'Alessando & Castellengo (1993), the pitch perceived for short vocal vibrato tones has been measured using a method of adjustment. The stimuli were synthetic vocal tones, produced by a formant synthesizer. The parameters under study were the tone duration, the frequency extent, the vibrato rate, and the nominal frequency as a function of the fractional number of vibrato cycles.

All the subjects noted that for several patterns it was possible to perceive a dynamic tone (a glissando, or gliding tone), rather than a static tone.

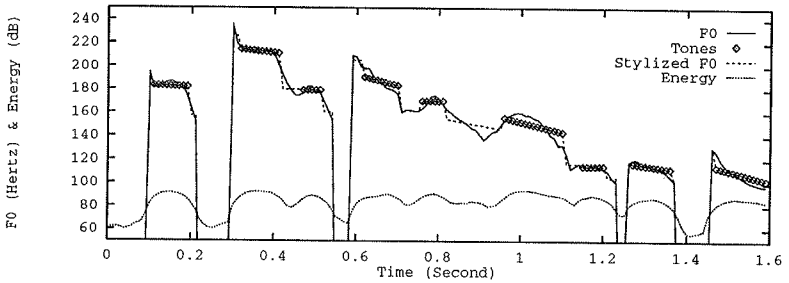

Figure 1: *Tones, FØ, stylized FØ and short-term energy.*

A dynamic tone is defined as a perceived tonal movement: the the FØ pattern (the physical variation of the fundamental frequency) results in a psychologically gliding tone (or glissando). For a static tone, the FØ pattern results in a single pitch percept.

A separate perception took place for the high and low parts of the vibrato cycles, for large frequency extents or slow vibrato rates (long duration). The same type of phenomenon was pointed out by Nabelek & al. (1970) in their study on pitch of tone bursts with changing frequencies. Our experimental data on fused (static tones) and separated (dynamic tones) pitch perception have been be consistently related to the glissando threshold.

Many psychoacoustic and psychophonetic data on the glissando threshold are available. One can find an interpolation procedure and a unified view of these data was presented in 'tHart and al. (1990). They studied the distribution of the glissando thresholds published in the literature and they showed that the glissando thresholds where distributed around a curve G_{tr} (expressed in Semi-Tones/Second) which approximately satisfies the equation:

$$\log(G_{tr}) \simeq -2.00 \times \log(T) - 1.83 \quad (1)$$

where T is the duration of the tone. 'tHart et al. reported that more than 75 % of the data in the literature lie within a distance of a factor of two from Equation 1, i.e. within the interval $[\log(G_{tr}) - \log(2), \log(G_{tr}) + \log(2)]$, in the double logarithmic scale. The

fusion/separation situations observed in our experiments are in good agreement with the glissando threshold. The glissando rates in case of separation are in the 75 % interval around the glissando threshold. In case of fusion, the glissando rates are all below the threshold.

FØ INTEGRATION

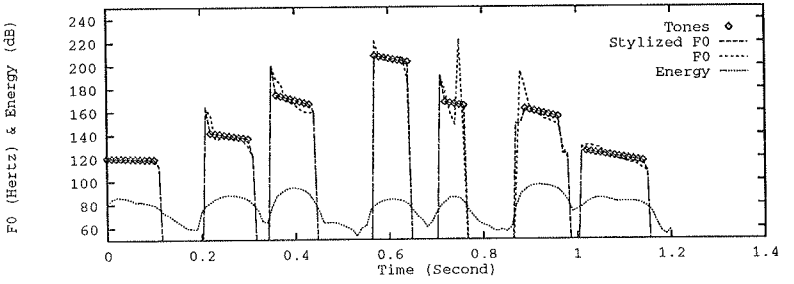

Figure 2: *Tones, FØ, stylized FØ and short-term energy.*

It appeared in the experiments that the final part of the tone had a larger weight on the pitch judgement than the initial one. The experimental results also suggested that the FØ patterns were time-averaged, at least in case of fusion. A quantitative model for such a process may be a time-average of the FØ pattern viewed through a data window. A simple model for the data window is a raised exponential memory function. Let $p(t)$ denote the pitch perceived at time t, f the time-varying FØ function, beginning at time 0, and let α, β be two constants, we have:

$$p(t) = \frac{\int_0^t (e^{-\alpha(t-\tau)} + \beta) f(\tau) d\tau}{\int_0^t (e^{-\alpha(t-\tau)} + \beta) d\tau} \tag{2}$$

In Equation 2, the constant β accounts for time averaging, and the constant α accounts for weighting of the past.

In the case of separation, the excitation patterns due to FØ extrema become more separated in time or in frequency, and these extrema are perceived as independant auditory events. It is reasonable to assume that if the glissando rate for a given condition is larger than the glissando threshold for the same condition, the amount of constant time-averaging represented by β is reduced. It does mean that in the separation case, the pitch judgement does not take into account the distant past. Therefore, Equation 2, with $\beta = 0$ is used when the glissando rate exceeds the glissando threshold.

The free parameters α and β have been estimated by minimizing the Root Mean Square distance between the model response and the experimental data. The optimal parameters obtained were $\alpha = -22$ and $\beta = 0.20$. The parameter β can be interpreted as the amount of long-term time-averaging (here $\beta = 20\%$). The parameter α represents the speed of decay of the exponential function. The bandwidth B of the corresponding low pass filter is: $B = \alpha/\pi = 7.00Hz$, and its time constant is $1/B = 0.14s$. Its seems that this optimal integration time constant can be consistently related to the auditory persistence.

APPLICATION

The model presented above was implemented in a computer program for tonal analysis. This progam contains several procedures : 1) pitch detection and voice/unvoiced decision. 2) short tone decomposition using short term energy. 3) glissando rate computation, and

static/dynamic tone decision. 4) for static tones, the perceived pitch value is computed, according to Equation 2. For dynamic tones, two values are computed, at the beginning and end of the tone, according to Equation 2 with $\beta = 0$. These two target values are linearly interpolated.

This system was tested on a corpus of read French speech. It appeared that only two broad types of tone were necessary and sufficient to describe the intonation (static tones, and unidirectional dynamic tones). In order to test this stylization procedure, we synthesized the unprocessed and stylized tonal contours using LPC synthesis. The natural and synthetic sentences were generally perceptually indistinguishable. Errors occured only when the segmentation procedure failed (i.e. the tonal decomposition was wrong).

Figure 1 shows the application of the tonal analysis to the sentence "cessez de faire du bruit les enfants!" [sesedøfɛrdybrɥilezãfã]. Figure 2 shows the sentence "ils sont partis pour Paris" [ilsõpartipurpari].

CONCLUSION

Further work is needed in order to understand the pitch of short tones. Particularly, no quantitative data and no model are yet available for describing the pitch of short-tones with joint time-varying energy and FØ patterns. Some work on glissando threshold in context is also needed : the glissando thresholds used in this study are probably too severe for continuous speech. It seems that the main problem remains tonal (or syllabic) decomposition.

The study presented herein concerned French, which is sometimes described as a "syllable-timed" language. For our corpus of read French, it appeared that only two types of short-tone patterns were sufficient: static tones and unidirectional dynamic tones. More complex patterns might be encountered in other languages or dialects (for instance in some dialects of popular French, rise+fall patterns can be found on a same syllable).

Our system describes intonation in terms of a succession of perceptually relevant tones. It provides a "tonal score" which can serve as a reliable basis for studying the melodic structures of speech.

REFERENCES

C. d'Alessandro & M. Castellengo, (1993). "The pitch of short-duration vibrato tones", preprint LIMSI, NDL 93-04, to appear in J. Acoust. Soc. Am.

D. House, (1990). *Tonal Perception in Speech*, Lund University Press.

P. Mertens, (1987). *L'intonation du Francais. De la description linguistique à la reconnaissance automatique.*. Unpublished doctoral dissertation, Catholic University of Leuven.

I. V. Nabelek, A. K. Nabelek & I. J. Hirsh (1970). "Pitch of tone bursts of changing frequency", J. Acoust. Soc. Am. **48**, 536-553.

J. 'tHart, R. Collier, and A. Cohen (1990). *A perceptual study of intonation*, (Cambridge Univ. Press, UK).

C. S. Watson, D. C. Foyle & G. R. Kidd (1990). "Limitted processing capacity for auditory pattern discrimination". Acoust. Soc. Am. **88**, 2631-2638.