PITCH CONTOUR STYLIZATION USING A TONAL PERCEPTION MODEL

P. Mertens and Ch. d'Alessandro
CCL, University of Leuven, Belgium and LIMSI, Orsay, France

ABSTRACT

An algorithm for automatic pitch contour stylization is described. It is based on a model of tonal perception, such that the resulting stylization is controlled by two perceptual thresholds. The output is the sequence of audible pitch events aligned with the syllables in the utterance.

INTRODUCTION

Stylization is a manual or automatic procedure that modifies the measured F0 contour of an utterance into a simplified but functionally equivalent form, i.e. preserving all melodic information which has a function in speech communication. There are several motivations for doing this; for instance, to reduce the amount of data required to generate pitch contours in synthetic speech, or to isolate the functional parts in the contour (and to remove the others) and to obtain a representation of this underlying contour, e.g. for intonation teaching or linguistic research. We propose an approach in which the stylization represents the pitch contour which is perceived by the average listener. In other words, the stylization process is seen as a simulation of tonal perception and as a way to measure what is heard. This approach satisfies both goals mentioned earlier: it results in an important data reduction and it filters out F0 events which cannot be heard and hence have no function in prosody. Still, there are other motivations for computing the perceived pitch: as a way to evaluate phonological intonation models and to obtain an automatic transcription of intonation. This may require some explanation.

There is little doubt about the acoustic manifestation of intonation (at least, for pitch): for most speech signals, F0 can be computed in an objective way, with estimation errors below perceptual thresholds. The phonological representation, however, is a sequence of symbols (tones, pitch movements) the determination of which involves a phonetician who interprets the data within a particular model. The large number of such models suggests the lack of a procedure to evaluate them. How could one decide that the descriptive units of model A are more viable than those of model B, or even that they are psychologically viable at all ? To this date, there is no clear criterion for the verification of intonation models; as a result their choice is often a matter of personal preference.

When someone describes the sounds he hears, he refers to the auditory image resulting from sensory and perceptual processing, rather than to the acoustic signal. It can easily be seen that the cognitive process of intonation understanding does not have direct access to the acoustic form (F0) but rather to the pitch events after processing by the peripheral auditory system and the perceptual system. Consequently it is this form that should be the input to the phonological model. By computing this perceived pitch contour, one will narrow the gap between the acoustic and the cognitive domain, because it eliminates one of the assumptions made by phonological models (namely the one about the nature of the input representation).

The rest of this paper gives a quick overview of tonal perception effects, then describes the algorithm implementing them. Finally we describe some of the results obtained.

Tonal perception

What is known about tonal perception?

1. Spectral and amplitude variations in the speech signal affect the way in which pitch variations are perceived, giving rise to a perceptual segmentation of the speech continuum [5]. This *segmentation effect* results in a sequence of short tonal events aligned with the syllables, rather than a continuous pitch curve for the whole utterance. Unfortunately, no quantitative model describing the contribution of changes in (global) amplitude, spectral energy, and other attributes, is yet available.

2. The perception of a changing pitch requires some minimal amount of frequency change as a function of time. Otherwise a static pitch is perceived. This effect is known as the *glissando threshold* (*G*). For a uniform pitch change (with constant slope), $G = 0.16 / T^2$, where T is the duration of the pitch variation. The effect has been investigated for years [3,4,8,9], both for pure tones and synthetic speech, but not in continuous speech.

3. A change in pitch slope will be perceived provided some minimal difference in slope, known as the *differential glissando threshold* (*DG*). There has been little research on this effect [4].

4. Static tones, i.e. short-term F0 variations which are below threshold G, are perceived with a certain pitch. In a study on the perception of vibrato [1], it was shown that this *short-term integration* can be modelled by a windowed time average (WTA) function.

Our stylization algorithm simulates these four effects.

DESCRIPTION OF THE STYLIZATION ALGORITHM

Figure 1 shows a block diagram of the stylization algorithm. It consists of several processing steps, some of which are purely acoustic (F0 measurement, voicing determination), while others are related to tonal perception. We will focus on the latter here. Most of the work in the algorithm concerns the determination of the speech fragments for which the perceptual effects are to be computed.

1. *Speech segmentation.* Since pitch perception is determined by spectral and amplitude changes, the speech signal is first divided into syllable-sized chunks. In the absence of a quantitative model of this effect, several types of segmentation are investigated. The first focusses on spectral change and uses the voiced parts of the syllables [2]; the second favours amplitude change and computes the syllabic nuclei (or loudness peaks) [6]. We will illustrate the results obtained with both segmentations.

2. *Short-term perceptual integration* of pitch. The WTA model is applied to the F0 in the voiced region of each syllable. This results in a smoothed pitch contour (as can be seen in Figure 2).

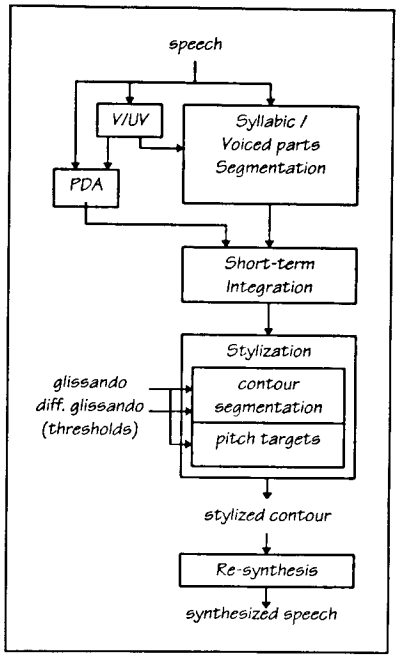

Figure 1. Block diagram of the pitch contour stylization algorithm. PDA is the pitch determination algorithm. V/UV is the voicing decision.

3. *Syllabic contour segmentation.* Syllabic pitch contours can be compound (e.g. rise-fall); they should be divided into simple, uniform parts first. This results in one or more *tonal segments* per syllable: a monotonous pitch change, i.e. either level, rising, or falling, and without an audible change in slope. This segmentation is motivated by the fact that the *G* and the *DG* are obtained for, and should therefore be applied to such uniform segments.

The syllabic contour segmentation involves two steps. The first locates the turning points in the contour so as to break it down into candidate tonal segments. The second makes a decision as to which candidate segments are to be grouped. The first step is recursive. Within an analysis interval with an audible pitch change (above *G*), a new turning point is found at the point of

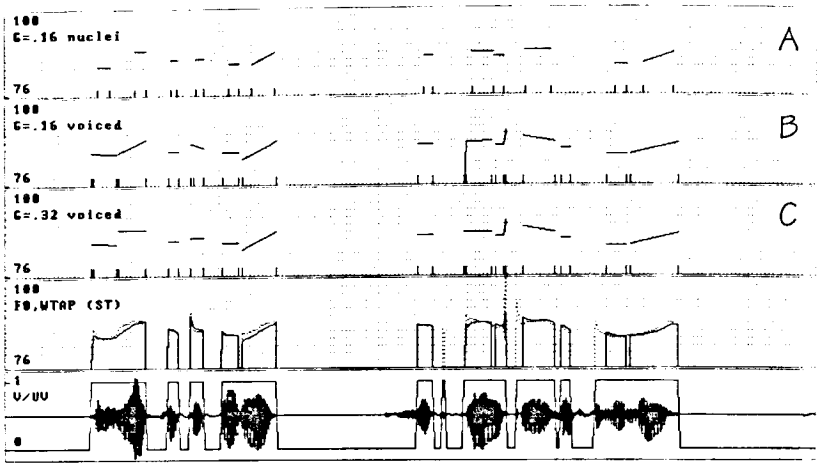

Figure 2. Comparison of stylization results with different parameter settings, for the utterance "d"ailleurs quand t'es pas là, [pause] tu sais de quoi on cause entre nanas", with a signal duration of 3.58s.. See text for explanation.

maximum difference between the observed (WTA-)pitch contour and a straight line between the start and end pitch, provided this difference exceeds some critical value of 1 ST. When a turning point is found the same procedure is applied recursively to both parts, before and after the turning point. The second step merges two consecutive candidate segments if their slope difference is smaller than the *DG*. It proceeds from left to right. After each merger the list of candidate segments is updated.

An interesting property of this procedure is that it is entirely controlled by two parameters, which are perceptual thresholds *G* and *DG*.

4. *Stylization*. For each tonal segment obtained in the previous step, the estimated *pitch targets* are obtained as the (WTA-)pitch at the start and end of the tonal segment. The stylized contour is the linear interpolation between successive pitch targets. For static tonal segments the pitch of the end point is extended to the entire tonal segment.

RESULTS

In order to evaluate the stylization, the speech signal was resynthesized (TD-PSOLA) with the stylized pitch contour and presented to 20 listeners for comparison with the original signal. The results of this experiment are described in [2]. By changing the two parameters of the model (*G* and *DG*) it is possible to evaluate their impact on the stylization. By a systematical evaluation of the parameter settings in listening tests the system can be used to determine the thresholds *G* and *DG* in continuous speech.

Figure 2 shows the stylizations obtained with different types of segmentation and different settings of the model parameters *G* and *DG*. The figure contains five parts. The lower part displays the speech signal together with the V/UV decision. All others parts use a semi-tone (ST) scale for the Y axis, with grid lines 2 ST apart. The next part shows the F0 (dotted line) and the WTA pitch (full line). The latter is calculated on the F0 values in the voiced part of the syllable. This results in a smoothed and somewhat time delayed version of the F0. The three upper parts show stylizations obtained with different parameter settings. The upper stylization (**A**) uses a segmentation into syllabic nuclei and the "standard" glissando threshold *G*=.16 (this value is the numerator in the equation given earlier).

The small vertical marks delimit the tonal segments. The two other stylizations (**B**,**C**) use the segmentation into voiced syllabic parts with parameter *G*=.16 or *G*=.32, respectively. By using *G*=.32 the glissando threshold is doubled, simulating the hypothesis that in continuous speech pitch changes twice as large as those for isolated stimuli would be required in order to be audible. All three stylizations are obtained with parameter *DG*=20, although a setting of *DG*=40 produces the same result. Stylization **B** gives the largest number of dynamic tones, while in **C**, due to the higher value of *G*, two dynamic syllables have been stylized as static ones. In **A** there are even less dynamic syllables because the nuclei are generally shorter than the voiced parts (which has an impact on the *G*). The second part of the utterance contains a F0 detection error, which is present as a dynamic tonal segment in **B**, and as a static one in **C**.

DISCUSSION

The stylization based on tonal perception has several inherent advantages over other types of stylizations.

1. It gives both a *qualitative* and a *quantitative representation* of the auditory contour, showing *how* the contour is perceived (which parts are perceived as dynamic, which as static, and which parts are not audible), and *what pitch* is perceived, for any time instant *t*. While many stylizations are descriptively adequate, ours also offers explanatory adequacy. In this respect, pitch movement approaches (e.g. "close-copy stylization") are less elegant because they sometimes suggest that the listener hears a changing pitch (in unvoiced syllable onsets, e.g.) where actually he hears no pitch at all.

2. It is *theory-independent*: it isn't linked to a particular prosodic model; it doesn't refer to pitch levels (which would have to be identified), to a declination line (which would have to be determined), to normalized pitch movements or contours, and so on.

3. The stylization proceeds from *left to right*, and can be applied to speech fragments as small as syllables. As a result one does not need the entire utterance to calculate the stylization (as is the case for declination line based procedures).

The stylization can be used as a tool for basic research.

1. By varying the model parameters, in combination with resynthesis and listening tasks, it can be used to measure the perceptual thresholds *G* and *DG* for continuous speech, at least for speech signals with an obvious segmentation into syllables.

2. The stylization provides an automatic transcription of the perceived tonal events, while eliminating the bias of the human transcriber. As such it is useful in the construction and verification of prosodic models.

REFERENCES

[1] Alessandro, C. d' & Castellengo, M. (1994), "The pitch of short-duraton vibrato tones", *J. Acoust. Soc. Am.* 95, pp. 1617-1630.

[2] Alessandro, C. d'; Mertens, P. (forthcoming, 1995), "Automatic pitch contour stylization using a model of tonal perception", *Computer Speech and Language*.

[3] Hart, J. 't (1976), "Psychoacoustic backgrounds of pitch contour stylisation", *I.P.O.- Annual Progress Report* 11, pp. 11-19.

[4] Hart, J. 't; Collier, R. & Cohen, A. (1990), *A perceptual study of intonation*, Cambridge: Cambridge Univ. Press, 227 pp.

[5] House, D. (1990), *Tonal Perception in Speech*, Lund: Lund Univ. Press

[6] Mertens, P. (1987), "Automatic segmentation of speech into syllables", in Laver, J. & Jack, M.A. (eds.), *Proc. of the European Conf. on Speech Technology*, vol. II, 9-12.

[7] Mertens, P. (1989), "Automatic recognition of intonation in French and Dutch", *Proc. Eurospeech* 89, vol 1, pp. 46-50.

[8] Rossi, M. (1971), "Le seuil de glissando ou seuil de perception des variations tonales pour la parole", *Phonetica* 23, pp. 1-33.

[9] Rossi, M. (1978), "La perception des glissandos descendants dans les contours prosodiques", *Phonetica* 35, pp. 11-40.