Automatic pitch contour stylization using a model of tonal perception

Christophe d'Alessandro* and Piet Mertens†

LIMSI-CNRS BP 133, F-91403 Orsay, France and †Department of Linguistics, K. U. Leuven BP33, B-3000 Leuven, Belgium

Abstract

A new quantitative model of tonal perception for continuous speech is described. The paper illustrates its ability for automatic stylization of pitch contours, with applications to prosodic analysis and speech synthesis in mind, and evaluates it in a perception experiment.

After a discussion of the psycho-acoustics of tonal perception, and an overview of existing tonal perception models and systems for automatic analysis of intonation, the model and its computer implementation are described in detail. It includes parameter extraction, segmentation into syllables, perceptual integration of short term pitch change, tonal segment computation, and pitch contour stylization.

This is followed by a perception experiment in which subjects are asked to distinguish original signals from resynthesized signals with automatically stylized pitch contours. The aim of this experiment is to show the usefulness of the model as a basis for intonation representation, and to study the influence of the model parameters. It is shown that the stylization obtained with the model is an economic representation of intonation which can be useful for speech synthesis and prosodic analysis. © 1995 Academic Press Limited

1. Introduction

Modelling the perception of intonation is a challenging problem to both fundamental and applied studies of prosody. A computer model of intonation perception should be able to process the acoustic speech signal, and to yield a quantitative representation of how the prosodic attributes of the signal are perceived. Perception research over the last 30 years has indeed shown the complex nature of auditory perception in general, and of tonal perception in particular. Recent work by House (1990), for instance, showed how spectral and amplitude variations create a perceptual segmentation of the signal into syllable-sized chunks. This process drastically influences the perception of prosodic attributes (such as pitch, stress and duration), since it transforms the continuous speech signal into a concatenation of short duration fragments. Earlier work established the glissando threshold, a perceptual threshold for fundamental frequency variation depending on the extent and the duration of the variation. Taken together, these two mechanisms can account for several observed phenomena. For instance, they explain

why many short-term fundamental frequency (F0) variations will go unnoticed. Along the same line, microprosodic variations may contribute to segmental perception but probably not to the perception of sentence intonation. All these facts need to be taken into account by the perceptual model.

The need for a perceptual model of intonation is felt in several areas, like (automatic) intonation analysis, (automatic) intonation transcription and intonation synthesis.

A key point in automatic analysis of speech intonation is that the perceived pitch is not always the same as the physical F0. The F0 pattern, as it appears at the output of a pitch tracker, seems difficult to interpret in a straightforward manner, for reasons which are well known. The F0 pattern depends upon several independent factors some of which are well described in the literature: e.g. intrinsic pitch of vowels and consonants, co-intrinsic pitch phenomena linked to the place and mode of articulation of the segments, voice source characteristics, loudness, phonatory force, etc. One can notice that a similar situation is encountered in singing. In many situations, the melody indicated by the vocal score, which is accurately appreciated by the audience, is rather different from F0 tracings.[1]

Automatic analysis in turn is a crucial step towards the long awaited automatic transcription of intonation. If this goal can be reached, it would make the work of prosodic transcription less tedious and cumbersome, and would give the transcription the objective basis that it is still lacking. Since the automatic analysis procedure applies the findings of psychoacoustic experiments, it provides a systematic transformation of the acoustic F0 data into an estimate of the perceived pitch, free from any bias introduced by the individual human transcriber. Automatic analysis and transcription will provide better tools for corpus analysis in phonetics and linguistics, because it allows for the gathering of large amounts of data. At the same time, it can serve as a stimulating test for modern theories of intonation in linguistics, phonetics and acoustics. Eventually it will narrow the gap between the acoustic data and the tonal decomposition assumed by linguistic analyses.

In the area of synthesis of prosody, the rationale behind the search for a perceptual model of intonation is the need to generate a natural intonation contour with a minimal amount of information. Rather than reproducing complete intonation contours, the goal is to find the perceptually relevant parts and properties of the contour and to generate the other parts from there.

In this paper, a new psychoacoustic model of pitch perception for short tones (d'Alessandro & Castellengo, 1994) is applied to the problem of automatic stylization for syllable-sized units. The aim is to compute one or several perceptually motivated tonal movements for each syllable. An important property of the model is that it is hoped to be language independent because it models perception prior to any process of linguistic categorization.

The next section presents the different steps that are needed in a perceptual model of intonation, emphasizing the problems addressed within the scope of the paper. It also gives an overview of existing systems for automatic analysis of intonation, from the point of view of intonation modelling, and presents the perceptual knowledge used in the stylization algorithm. Section 3 describes an algorithm for automatic intonation

---

[1] Observing F0 tracings in singing, Seashore noted as early as 1938 that "It is shockingly evident that the musical ear which hears the tones indicated by the conventional notes is extremely generous and operates in the interpretative mood" (Seashore, 1938, p. 269).

stylization. The procedure for automatic stylization was tested using a formal perception experiment based on resynthesis of stylized contours. This experiment is described in Section 4. The final section summarizes the work done, discusses the results obtained, and indicates some future developments.

2. Towards a perceptual model of intonation

2.1. A review of automatic analysis of intonation

A variety of techniques and methods inherited from pattern matching, perception modelling, expert systems, and natural language processing have been applied to automatic analysis of intonation. Automatic analysis of intonation comprises several distinct aspects, including automatic F0 stylization, stress detection, recognition and classification of intonation units, prosodic parsing.[2]

Intonation stylization is a specific problem, which is different from the problem of prosodic recognition. The aim is not to identify linguistic features, but to retain only those parts of the F0 contours which are perceptually relevant. Therefore, intonation stylization should involve some sort of resynthesis. This aspect was also studied in a number of works.

Studies on automatic F0 stylization can be divided into two major groups, according to the presence or absence of perception modelling. A first group, using an acoustic approach, often uses linear regression analysis to obtain a stylized F0 curve. Whenever the correlation between successive F0 values drops below a fixed value, a boundary between successive line segments is found (e.g. Kloker, 1976; Rietveld, 1984). Huber (1990) used the same technique to construct grid lines, by computing the correlation between successive peak values, or between successive trough values, generating a change of grid at those points where the correlation drops.

Only a few studies incorporate knowledge about tonal perception in at least some part of the analysis and stylization process. We take a closer look at them here. For brevity parameter extraction (F0, intensity, voicing, etc.) is not discussed.

The idea to stylize pitch contours originates from the research conducted in the mid sixties at the Institute of Perception Research of Eindhoven (IPO). Pitch contour stylization is based on the assumption that the pitch contour of an utterance can adequately be synthesized, and hence be represented, by a sequence of straight lines. This representation is obtained via an interactive procedure, known as "close copy stylization", in which a subject judges the resynthesized utterance for which the F0 data have been replaced by a curve consisting of straight lines between points selected by the subject. In later studies, the straight lines obtained using this ad hoc stylization are replaced by a set of prototype straight lines, the "standardized pitch movements". Standard pitch movements constitute the basic intonation units for the language under consideration. The original model for Dutch has been adapted to other languages (t'Hart, Collier & Cohen, 1990). The timing specifications of the standard pitch movements refer to vowel onset positions, so these have to be supplied. Implicitly the timings also refer to syllable durations for a normal speech rate. The lack of a

---

[2] The following studies were devoted to automatic recognition of prosody (i.e. stress and intonation units detection, prosodic parsing): Lea, Medress and Skinner (1975), Kloker (1976), Rietveld (1984), Gibbon and Braun (1988), Waibel (1988), Vaissière (1988), Huber (1990), Wightman and Ostendorf (1991, 1992), Geoffrois (1993), Carbonell and Laprie (1993), Bagshaw (1993).

preliminary segmentation into syllables makes the model of course very economical, but less perceptually justified (see also Kohler, 1991, pp. 121) for a criticism of this approach). Several attempts towards automatic straight line stylization have been made (t'Hart, 1979; ten Bosch, 1993). This type of stylization is based on F0 alone, and not on intonation perception. Formal assessment of the stylization process has been carefully conducted for several languages. Section 4 reports some comparison of the results obtained using the IPO methodology and the automatic stylization procedure proposed below.

Rossi, Di Cristo, Hirst, Martin and Nishinuma (1981) give a good overview of the work on stylization completed at the Institut de Phonétique d'Aix-en-Provence (IPA), for a variety of languages. The proposed stylization process takes advantage of perceptual thresholds. A detailed description can be found in Nishinuma (1979, pp. 109–121), presenting a prosodic analysis system for Japanese. In contrast to the IPO approach, Hirst, Nicolas and Espesser (1991) proposed a stylization procedure in terms of target values, rather than straight lines, for fully sonorant segments. Natural and synthetic contours are compared in terms of visual similarity. No formal perceptual comparison on the quality of intonation stylization is reported.

In his study on French, Mertens (1987a) distinguishes three levels of representation: the acoustic, the perceptual, and the linguistic levels, and hence two main processing blocks: the simulation of perception and intonation parsing. The former block can be seen as a quantitative model of tonal perception and the whole as a system for automatic prosodic recognition. Mertens (1989) updates the system and extends it to Dutch intonation. The system includes the following processing steps:(1) segmentation into syllabic nuclei and pauses (cf. Mertens (1987b), the nucleus corresponds to the high energy voiced part of the syllable); (2) normalization of co-intrinsic microprosody (at vowel onset) and detection of pitch extraction errors; (3) simulation of perceptual processing for each syllabic nucleus (glissando threshold, dependence of loudness upon duration); (4) categorization of syllabic prosodic features (such as static/dynamic, slope type, long/short, stressed/unstressed, stress type); the stress detection (from loudness and duration) uses heuristic rules; and (5) phrase-structure parsing of the list of syllables (with their partially determined features) according to a grammar of intonation for the target language. The last part transforms the system into an understanding system, which recognizes prosodic forms and assigns a structural interpretation to them on the basis of their distributional (i.e. syntactic) properties. Stylization is seen as a side-effect of the perceptual integration and of the model-specific categorization. No resynthesis experiments are reported.

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 (pp. 108–118), along with an automatic F0 stylization procedure. The results on the influence of spectral changes on intonation perception serve as a basis of the stylization procedure proposed below. In the Lund experiments, intonation stylization is performed by linear (straight-line) interpolation between target values (the averaged F0 values in a 32 ms window) at vowel onset and at the end of the syllable. LPC resynthesis of utterances with their stylized F0 curve indicates that the majority of the stylized sentences could not be distinguished from their original counterparts, although the reduction did give rise to a few cases of clearly audible tonal deviation. It must be emphasized that this statement was based on

informal listening: it will be discussed in the light of our formal experiments in Section 4.

All intonation analysis systems incorporating some sort of perceptual modelling have two things in common. (1) They acknowledge the need for an initial segmentation into syllable-sized units (syllables, syllabic nuclei, implicit vowel onsets for IPO). However, the segmentation itself is not done on perceptual grounds, it is either manual or based on acoustic properties. (2) Heuristic rather than perceptual rules are used for pitch contour stylization, because at the time these systems were built too little was known about tonal perception. The most advanced systems clearly separate perceptual integration (of pitch and other prosodic attributes) from linguistic categorization (of intonation units). Data on the quality of intonation stylization, based on formal testing, is available only in IPO work.

2.2. *The components of a perceptual model of intonation*

In summary, the following levels of processing can be identified in a perceptual model of intonation.

(1) Parameter determination (e.g. F0, energy, zero-crossing rate, voiced–unvoiced (V/UV) detection, ...).

(2) Primary segmentation of the speech signal, and hence of the pitch contour, into syllable-sized segments. This primary segmentation is based primarily upon relatively large spectral variations and overall energy variations.

(3) Perceptual integration of short-term intrasyllabic pitch variation, or perceptual F0 smoothing.

(4) Secondary segmentation of the integrated pitch contour into successive tonal segments (these terms will be defined later on) and perceptual integration of mid-term pitch variation, according to the glissando threshold (static vs. dynamic tones) and the differential glissando threshold (rise–fall, rise–rise combinations, etc.). At this stage of processing, pitch targets (or pitch movements) are assigned to the tonal segments.

(5) Categorization of the tonal segments within the language specific prosodic categories. Here, the various sources of information (durations, pitch targets, loudness) are combined and interpreted in order to map the perceptual events onto the linguistic units. For instance the micro-prosodic aspects will be separated from the intonational aspects at this stage of processing. Also stress will be determined on the basis of duration, loudness, and pitch data.

These processing steps are all strictly local, i.e. restricted to a single syllable, and are followed by non-local processing in the intonation parser, which takes into account several syllables.

The model developed in this paper will address only points (1)–(4) of the program presented above: parameter extraction, primary syllabic segmentation, perceptual integration and the secondary segmentation of the integrated pitch contour.

Three basic properties of tonal perception are exploited in the processing scheme proposed above, and have to be examined here. The first property is the primary segmentation of intonation into syllable-sized chunks. The second property is related to F0 integration. The third property is linked to absolute and differential thresholds of pitch change (glissando and differential glissando thresholds).

2.3. Intonation segmentation

It is generally held that a syllabic decomposition takes place at an early stage in speech perception. It is reasonable to assume that a model of intonation perception should be based on the segmentation introduced by the syllabic stream.

Following Kohler (1991, p. 122), we assume that F0 contours should not be interpreted in isolation, but rather in conjunction with other phonetic and prosodic characteristics of the speech signal. For instance there is some evidence (House, 1990, pp. 36–63) that the same F0 movement will be perceived either as a pitch glide or as two level tones depending on the segmental context, because of the inherent perceptual segmentation of the speech signal into syllabic units due to spectral and intensity change.

This perceptual mechanism transforms the utterance's F0 contours into sequences of short-duration tones. No quantitative data are yet available on this segmentation mechanism, and more work is clearly needed. In the following, syllabic segmentation will be considered as a first approximation of the tonal contour segmentation process.

2.4. F0 integration and the WTA model

Fundamental frequency is a physical parameter: it should not be confused with the perceived tonal height, or pitch. Since the pitch will be estimated from F0, the accuracy of the tonal perception model is in principle limited by the accuracy of the initial F0 measurement. Under good conditions, for long pure tones, the difference limen (DL) for F0 changes is around 0·3 to 0·5% (Hess, 1983, p. 78). These DLs can degrade by one order of magnitude for short tones extracted from natural speech. According to a comparison of the accuracy of pitch trackers with that of humans (Hess, 1983), it appears that the former are accurate enough to assume that the measurement of F0 contour introduces, on the average, no or very little error. Pitch determination algorithms (PDAs) can introduce a smoothing of the F0 data, due to the analysis window (typically 25 to 40 ms); this is the case for frequency domain PDAs, in particular.

The auditory system seems unable to follow rapid changes in fundamental frequency. There is some evidence that an integration process takes place in pitch perception for short-term F0 changes. d'Alessandro and Castellengo (1994) demonstrated this integration phenomenon in a study on vibrato perception. They proposed a weighted time-average model (WTAM) for pitch perception of short tones with time-varying fundamental frequency. In their study, a large set of experimental data on pitch perception for short-duration tones with changing frequency were obtained. It appeared from 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 F0 patterns were time-averaged. A quantitative model for such a process has been proposed. It consists of a time-average of the F0 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 F0 function, beginning at time 0, and let α be a constant. The WTAM for pitch perception then is:

$$p(t) = \frac{\int_0^t e^{-\alpha(t-\tau)} f(\tau)\, d\tau}{\int_0^t e^{-\alpha(t-\tau)}\, d\tau} \tag{1}$$

where the constant α accounts for weighting of the past. The free parameter α has been

estimated by minimizing the Root Mean Square distance between model response and experimental data. The optimal value was found to be $\alpha = 22$.[3]

This function represents the actually perceived pitch, i.e. the fundamental frequency of a static tone that would give rise to the same pitch judgement at the considered point in the pitch contour.

We think that the model obtained for singing is fully applicable to intonation analysis in speech, because: (1) the durations, extents and F0 patterns used in d'Alessandro and Castellengo (1994) experiments are comparable to those observed in speech; (2) the psychological thresholds are probably more severe for musical perception, compared to speech perception; (3) these thresholds are also probably more severe for short-tones in isolation, compared to short-tones in context (Watson, Foyle & Kidd, 1990).

However, a crucial difference between the listening conditions in the vibrato experiment and those of continuous speech is that in the former the listener was asked to judge a pair of isolated stimuli repeated several times, whereas in the latter, the tone appears in the context of other tones and is only given once.

In the context of prosody, the only study on pitch integration in short tones that we have been able to locate is in the work of Rossi (1971, 1978). He postulated the so-called perceptual "2/3 rule", which can be stated as follows:

> For dynamic tones in a vowel, the pitch perceived corresponds to a point between the second and the third third of the vowel.

The WTAM applied to unidirectional linear frequency glissandi is fully in agreement with Rossi's rule. Compared to Rossi's result, the WTAM is able to predict more accurately where is the "point between the second and the third third of the vowel". Furthermore, it is able to predict the pitch perceived for any F0 contours, and not only unidirectional glissandi.

2.5. Audibility of pitch changes

2.5.1. Tones, tonal segments, and pitch targets

Some terminology will be introduced here in order to avoid confusion. In our case, a *tone* is the pitch object perceived for a stretch of speech corresponding to a phonetic syllable (as defined below). The tone can be either *static* or *dynamic* depending on the absence or presence of a perceived pitch movement within the syllable. Within a given tone one or more *tonal segments* may be isolated, for which the perceived pitch shows a uniform slope (either level, rising or falling). Whereas static tones have just one tonal segment, dynamic tones may have more. A *simple tone* contains exactly one tonal segment (either level, rise or fall). A *compound tone*, or *complex tone*, is a combination of two or more tonal segments (e.g. in concave or convex tones). It is hypothesized that any pitch contour may be represented by a concatenation of tones, and therefore of tonal segments, both at the syllable and at the utterance level.

Since each tonal segment has a uniform slope, it can be represented by one or two *pitch targets* (for simple static tones and simple dynamic tones respectively). Put the

---

[3] In the work by d'Alessandro and Castellengo, two equations have been proposed for the WTAM, depending on the F0 extent of the stimuli compared to the glissando threshold. In the present work we take only the form of the WTAM which corresponds to the situation where the F0 extent of the stimuli are above the glissando threshold.

other way round: if more than two targets are needed, then we are dealing with a compound tone, consisting of two or more tonal segments, and there must be a change in slope at the boundary between two consecutive tonal segments.

For syllable pitch contours, which are short-duration tones, a model of tonal perception should determine whether they are simple or compound tones. In the latter case a segmentation of the syllable pitch contour has to be made. In the former case, the tone contains only one tonal segment.

A model of tonal perception should also tell us which pitch is perceived for every point in the contour. The question of the perceived targets involves a numerical model of pitch perception.

The aspect of the audibility of pitch changes is related to the glissando threshold, which has been studied in many psycho-acoustic experiments. The aspects of the segmentation of complex tones and of the perceived pitch have received very little attention in the literature. These three aspects are discussed in the next sections.

But let us first add a note on the distinction between the perceptual targets and those used in linguistic analyses of prosody.

In some syllables a sequence of tonal segments is heard, for instance when the pitch first rises (segment 1) and falls afterwards (segment 2). The maximum number of concatenated tone segments to be expected is limited by language-specific properties, in particular by the inventory of syllable pitch contours for the language.

In many phonological models of prosody the set of syllabic contours is represented as a sequence of discrete, static parts (or structural positions, or morae), for which target values (e.g. pitch levels) are used. The number of targets will depend on the set of contours to be identified. For French, it is generally acknowledged that the (linguistically distinct) syllabic pitch contours can be represented by two morae. This abstract representation used in linguistics should not be understood to imply an equivalence relation between the perceived tonal segments and the structural positions, but merely as an economic and functional reanalysis of the observed perceptual objects. As a result the number of perceived tonal segments can be larger than the number identified in the linguistic analysis. The perceived pitch contours can indeed consist of simple tones (such as static level, or dynamic rise and fall) or compound tones (such as late rise, late fall, rise-fall, etc.). Whereas a simple rising tone (two targets) and a late rise compound tone (three targets) differ at the perceptual level, they may both be represented by two morae by a specific linguistic analysis in which they would be considered as free variants of a single abstract intonation unit.

2.5.2. The glissando threshold

The perception of pitch for tones with changing frequency has been studied for years, particularly in the field of prosody. These studies generally focused on the audibility of pitch changes, which is related to the absolute threshold of pitch change, or *glissando threshold*. A *glissando* is an audible pitch change. Psycho-acoustic and psycho-phonetic data on the glissando threshold have been obtained by Sergeant and Harris (1962), Klatt (1973), Pollack (1968), Rossi (1971, 1978), and Schouten (1985).

The threshold varies with stimulus duration. The rate of frequency change over time is referred to as the *glissando rate*. The glissando rate specifies the slope of the frequency change.

A unified view of these data has been presented first by t'Hart (1976), and has been recently revisited by t'Hart *et al.* (1990).

The semi-tone per second ratio (ST/s) was proposed as the optimal unit for dealing with the glissando threshold. When using the frequency intervals on a logarithmic scale, a threshold is obtained, which is almost independent on the absolute frequency. Studying the distribution of the glissando thresholds published in the literature, t'Hart and co-workers showed that the glissando thresholds were distributed around a curve G_{tr} which approximately satisfies the equation:

$$G_{tr} = \frac{0 \cdot 16}{T^2} \tag{2}$$

where T is the duration of the tone, and where G_{tr} is expressed in ST/s. If Equation (2) is plotted using a double natural logarithmic scale, it becomes approximately a straight line, in a domain of variation for T which is compatible with syllabic durations (roughly: $T \in [0 \cdot 05s, 0 \cdot 200s]$):

$$\log(G_{tr}) = -2 \cdot 00 \times \log(T) - 1 \cdot 83 \tag{3}$$

t'Hart *et al.* reported that more than 75% of the data in the literature lie within a distance of a factor of two from Equation (2), i.e. within the interval $[\log(G_{tr}) - \log(2), \log(G_{tr}) + \log(2)]$, in the double logarithmic scale. The aim of most of the studies mentioned above is to estimate a psychological threshold. Therefore, the glissando threshold is generally obtained without any pitch measurement (except in the work of Rossi (1971, 1978)).

In most psycho-acoustic studies, only short tones without important spectral or energy variation are considered.[4] Of course this is not realistic for actual speech. However, due to the lack of a quantitative perceptual model for the interaction between pitch, intensity and spectral changes, it will be assumed here that the perceptual thresholds found for sounds without major spectral or amplitude change can be applied to the voiced parts of phonetic syllables as well.

For each tonal segment the perceptual model makes a static/dynamic tone decision. This decision is made on the basis of a comparison of the glissando rate for each tonal segment with the glissando threshold. The tonal segment will be labelled dynamic if it exceeds the threshold, and static otherwise.

2.5.3. Differential threshold of pitch change

t'Hart (1976, p. 17) studied how temporal proximity of tones affects our ability to distinguish their sizes or slopes. He stated that:

> It then turns out that, within the range of excursions that can be found in normal speech, more than one slope can be distinguished only for the longer durations.

---

[4] Clearly, tonal glides induce spectral changes, because of the variations in the excitation source of speech signals. But the spectral envelope, or "filter" part of the speech signal, is supposed invariant in most psycho-acoustic studies, where sustained vowels or pure tones are generally used.

The differential threshold of pitch change is the minimum difference in slope necessary to distinguish between two successive glissandi, more precisely between two successive tonal segments. It can be formulated in several ways.

Pollack (1968) and t'Hart *et al.* (1990, p. 33) use the ratio $g1/g2$, where $g1$ and $g2$ are the slopes of the first and the second part, respectively, each expressed in Hz/s. The required minimum ratio was found to be about 2 to 10, depending on the duration of the stimulus and on the slope value itself. Because by using a logarithmic scale (such as the semitone scale) thresholds and pitch intervals can be expressed without reference to absolute frequency, we will use the logarithm of the ratio $g2/g1$. With $g1$ and $g2$ expressed in Hz/s, we take the difference:

$$12 \times (\text{sign}(g2) \log_2(|g2|) - \text{sign}(g1) \log_2(|g1|)) \qquad (4)$$

which gives $g2 - g1$ expressed in ST/s. We propose to define the threshold as $g2 - g1$, with $g1$ and $g2$ expressed in ST/s, because this allows for a uniform treatment of arbitrary slope combinations, independently of slope direction. The difference has some interesting properties. On the one hand, the slope difference is positive for convex or positive slopes, and negative for concave or negative slopes. On the other hand, the magnitude is proportional to the amount of change, independently from the direction of the slopes. Using this convention, the differential threshold of pitch change was found to be about 12 to 40. Only a few experiments have been reported in the literature aiming at establishing the differential threshold, and it remains unclear whether their results can be applied to the prosody of speech. Nevertheless, the values found in the literature will be used in our experiments. -

3. Automatic stylization algorithm

The algorithm described here mimics the principles of speech prosody perception proposed above, yielding a perceptually based stylization of the F0 contour. An overview of the implementation of the model is presented in Fig. 1. Some details on this implementation follow.

3.1. Phonetic segmentation and syllabification

The model presented above assumes a syllabic segmentation of the tonal stream, which can be obtained directly from a phonetic segmentation. When the stylization is to be used in speech recognition or in automatic transcription of intonation, the input text is unknown and the syllabic segmentation should be text-free. Then, an open-set phone recognizer is appropriate.

Phonetic syllabic segmentation itself is beyond the scope of this paper. The assessment of the intonation model should be independent of segmentation errors. For this reason, we manually verified the output of an automatic phonetic recognizer. Our formal experiments are based on the timing information obtained from the LIMSI speaker-independent phonetic speech recognition system (Lamel & Gauvain, 1993). The resulting segmentation for the test corpora is accurate enough for the purpose of this research. In this case, the text corresponding to the speech samples was known, and we used the speech recognition system for accurate phone transcription. Furthermore, as we are interested only in syllabic decomposition, most of the few errors made by the speech
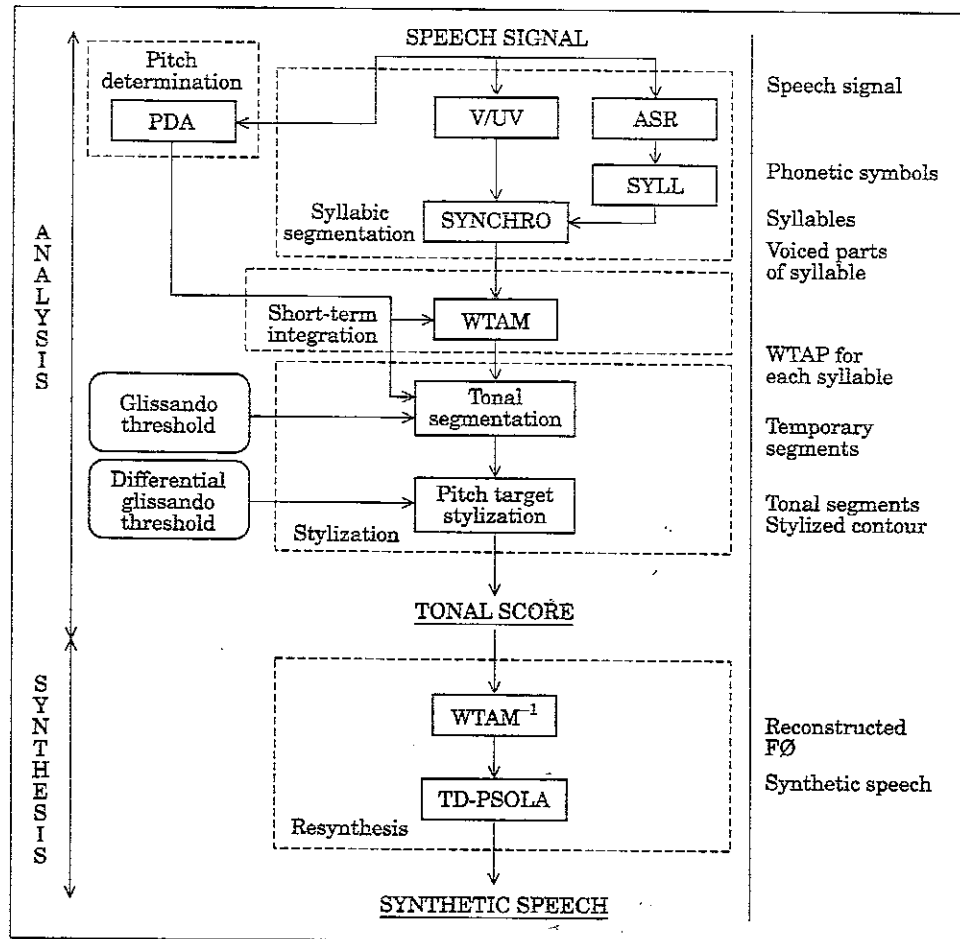
Figure 1. Overview of the system for automatic analysis of intonation. PDA:
pitch determination algorithm; V/UV: voiced–unvoiced decision; ASR:
automatic speech recognizer; SYLL: syllabation rules; SYNCHRO:
synchronization of voicing and phonemic boundaries; WTAM: weighted time-
average model; TP-PSOLA: time-domain pitch-synchronous overlap add.

recognition system do not introduce syllabification errors, but are phonetic labelling
errors (e.g. confusion between vowels, between nasal and voiced fricatives, etc.).

In the context of this study we will take the *phonetic syllable* to be a continuous
voiced segment of speech organized around one local loudness peak, and possibly
preceded and/or followed by voiceless segments. For instance the French word "so-
cialisme" can be pronounced as [sosjalismə] or [sosjalism] in which case the phonetic
syllables will be [so sja lis mə] and [so sja lis m] respectively. In the latter case the nasal
[m] forms a phonetic syllable by itself. Also note that [sjalis] is analysed as two syllables
because of the two loudness peaks, corresponding to the two vowels. For the purpose
of this paper phonological issues such as the internal structure of the syllable and
ambisyllabic consonants will not be considered. It is not obvious which syllable the [l]
belongs to, and whether the part of the pitch contour corresponding to the [l] should

be included in the previous phonetic syllable, in the next, or in both. Ambisyllabic consonants are of course located at syllable boundaries, where pitch perception is less accurate due to the spectral and amplitude changes (e.g. House, 1990). Ambisyllabicity may introduce an ambiguity in the syllabic decomposition process: the decomposition may not be unique. Therefore, ambisyllabicity must be handled consistently. If it is the case, it should only slightly affect the stylization process.

Phonetic syllables are derived from the phonemic segmentation via a small set of rules. As a result a set of phonetic syllables is available, together with F0 intensity and V/UV parameters. Since the V/UV and segmental timing data do not necessarily coincide, it was decided to give priority to the V/UV measurement over the segmental information. This was done automatically, by moving the syllables boundaries delivered by the recognizer, in order to align voiced segment boundaries and V/UV data.

When a syllable contains unvoiced parts, the tonal segment associated with it will coincide with the voiced part of the syllable. This guarantees that the concatenation of all the tonal segments preserves the original F0 and V/UV measurements.

3.2. Pitch determination and integration

In our experiments, F0 was determined every 10 ms using a modified version of the spectral comb method proposed by Martin (1982). Sampling frequency was 16 kHz, with a 37·5 ms Hamming window, 1333 Hz low-pass filtering, 256 points FFT (on the low-pass filtered signal), 1200 point cubic spline interpolation, frequency damping of 0·125 (for an explanation of the meaning of these parameters, see the description of this algorithm in Martin (1982)).

It must be noticed that the 37·5 ms frame introduces a smoothing effect of the F0 contour, particularly for high pitched voices. The time constant for this smoothing effect is about four times lower than the time constant of the WTAM (which is about 140 ms, see d'Alessandro & Castellengo, 1994). Post-processing of the pitch determination algorithm is based on a dynamic programming algorithm. For each frame, the spectral comb method delivers a vector of pitch candidates, each with a frequency and an amplitude. All the pitch candidates are kept for each analysis frame, and the optimal path among the frames (i.e. the path which cumulated the higher amplitude in the array whose lines were frame and columns were frequencies) is computed using a dynamic programming algorithm. This algorithm is applied to the whole set of frames for each sentence.

This post-processing proved good enough to remove all the octave errors introduced by the pitch detection algorithm, at least for the data in our test corpus. The spectral comb pitch detection algorithm has been recently revisited in the work by Hermes (1988) (Sub-Harmonic Summation method), where a discussion of the perceptual grounds of this type of frequency domain pitch determination algorithm can be found.

The voicing decision is taken on the basis of zero-crossing and energy thresholds. Isolated unvoiced or voiced frames are avoided by a post-processing step.

Fig. 2 shows an example of the F0 contour and of the phonetic and syllabic segmentation obtained at this stage of the algorithm. The short bars indicate the phoneme boundaries delivered by the speech recognition system, and the longer bars with black bullets indicate the syllable nucleus onsets, computed by rules. The x axis represents time (expressed in seconds), and the y axis represents frequency. Frequency is represented on a semi-tone (ST) logarithmic scale. The 1 Hz frequency is taken as

largest difference is selected as the turning point, and hence as a potential tonal segment boundary.

The segmentation is done recursively, starting with the entire syllabic pitch contour, splitting the analysis window at each turning point, unless a boundary condition is reached.

Two conditions will bring the recursive segmentation to a halt. (1) When the overall glissando rate of the analysis window is below the glissando threshold. Since in this case the observed pitch change is not audible, it is a static segment, and there is no reason to divide it into smaller parts. (2) When the pitch difference at the turning point (potential tonal segment boundary) is below some critical value, set to 1 ST. The latter condition avoids dividing a uniform contour part.

The recursive segmentation provides us with a list of monotonous contour segments. Each of them is a potential tonal segment, but not necessarily so. Any two contiguous contour segments will be grouped together, if the difference in slope is below the differential glissando threshold. This second pass proceeds in a left-to-right fashion, and updates the list of temporary segments each time two segments are merged.

As a result, the contour segmentation algorithm is based exclusively on two perceptual thresholds: the glissando threshold and the differential glissando threshold. Fig. 3 illustrates the segmentation of syllabic pitch contours into tonal segments.

3.3.2. Assignment of perceived pitch targets and stylization

Once the contour has been segmented, target values can be assigned. This essentially reduces to selecting values of the WTAP at the boundaries of tonal segments. For static tonal segments, the targets are set to the WTAP value at the end of the segment. Indeed, since no pitch change is perceived for these static segments, the WTAP at the end is the best available estimate of the perceived pitch. For rising and falling segments, the two target values are the WTAP values at the beginning and the end of the tonal segment.

The stylized pitch contour of a phonetic syllable is obtained by linearly interpolating (on a linear Hz scale) between successive pitch targets (of the voiced part, of course). There is some evidence that the choice of an interpolation function for pitch contours approximation is not psychologically critical (t'Hart, 1991). The stylized contour of an utterance is obtained by applying the same procedure to all syllables.

The algorithm described above can be useful for displaying diagrams that represent the perceived prosody. One of the aims of this algorithm is to give a perceptually motivated visualization tool, which is thought useful for prosodic analysis.

The representation of stylized pitch contours, together with syllabic marks will be called "Tonal Score". Fig. 4 gives an example of tonal score, obtained with the data of Fig. 2. It is rather easy to "read" intonation using the tonal score. Rhythmic information is provided by the string of syllable marks. These marks are located at syllable nucleus onsets, which are fairly good approximations of the perceptual centres of syllables. Melodic information is provided by the position of tones, in relation to rhythmic marks.

An illustration of the stylization for spontaneous speech is shown in Fig. 7 (F0 and corresponding tonal score). One can notice some of the particular features of spontaneous speech (compared to read speech), such as hesitations, pauses, large speaking rate variation within a same utterance.
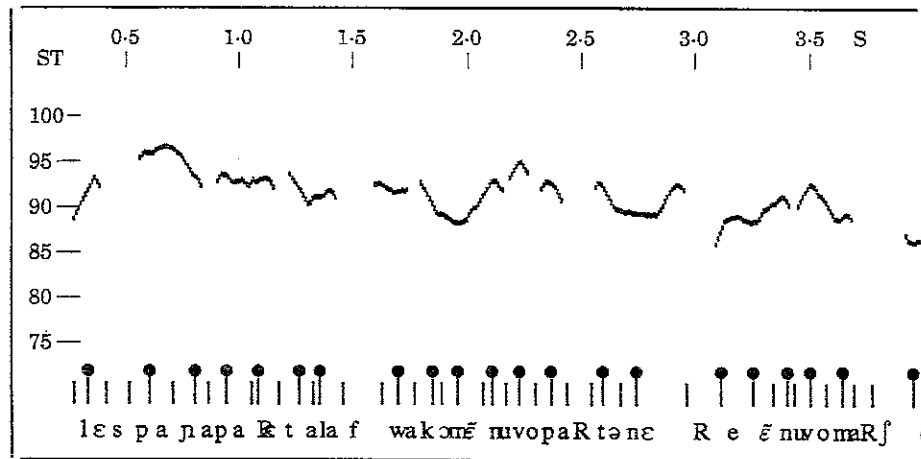
Figure 2. F0 contour and syllabic segmentation for the sentence: "L'Espagne apparaît à la fois comme un nouveau partenaire, et un nouveau marché". Female speaker.

reference. Then the frequency in ST is obtained as $12 \times \log_2(f)$, with f expressed in Hz. Using this convention, the frequency range of speech intonation, 80–500 Hz, corresponds to 75·86–107·59 ST.

The WTAM performs a linear smoothing of the data delivered by an F0 tracker, according to Equation (1). It is a passive model, which is supposed to account for the integration characteristics of the auditory system. A new integrated pitch contour, the Weighted Time Averaged Pitch (WTAP) contour is obtained after this stage.

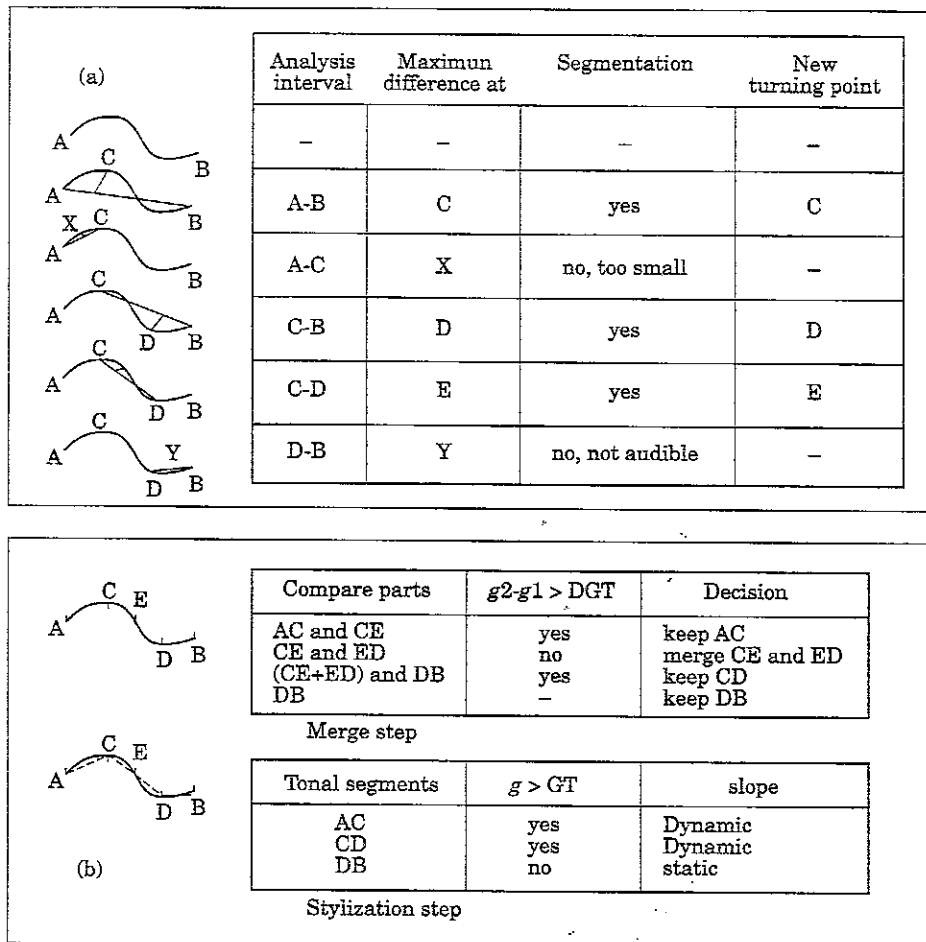### 3.3. Stylization of syllabic pitch contours

Before we can determine whether a pitch change is a glissando, i.e. is audible, we first have to delimit the time slice of the contour on which the glissando rate has to be computed. In many cases it will coincide with the entire syllabic contour, but for complex tones (e.g. rise–fall), all simple tonal segments have to be identified first.

As a result the stylization algorithm consists of two major steps. First the syllabic pitch contour is decomposed into a sequence of tonal segments. This is done on the basis of two criteria: the differential glissando threshold, and the glissando threshold. This step is followed by the actual stylization, in which pitch targets are assigned to each tonal segment.

3.3.1. Segmentation of compound tones

The basic idea underlying the segmentation of the syllabic pitch contour looks quite straightforward. We will locate the important changes in the contour, and break it up at those turning points.

Turning points in the WTAP contour are located by fitting a straight line between the WTAP value at the start and the end of a time window, and by evaluating the difference between the fitted line and the observed WTAP values. The point with the

	Analysis interval	Maximun difference at	Segmentation	New turning point
(a)	—	—	—	—
	A-B	C	yes	C
	A-C	X	no, too small	—
	C-B	D	yes	D
	C-D	E	yes	E
	D-B	Y	no, not audible	—

Compare parts	$g2\text{-}g1 > DGT$	Decision
AC and CE	yes	keep AC
CE and ED	no	merge CE and ED
(CE+ED) and DB	yes	keep CD
DB	—	keep DB

Merge step

Tonal segments	$g > GT$	slope
AC	yes	Dynamic
CD	yes	Dynamic
DB	no	static

Stylization step

Figure 3. Automatic stylization algorithm. A: Illustration of the recursive pitch contour segmentation, for the input data shown on top. A straight line is fitted between the pitch values at the start and end of the analysis interval. A turning point can be found at the point of maximum difference between the pitch data and the fitted line. See text for more details. B: Illustration of the merge step and the final stylization step, for the contour shown on top (dotted line). The merge step uses the differential glissando threshold (DGT) $(g2 - g1)$. The stylization step uses the glissando threshold (DG). The resulting stylization is shown as the dotted line on the lower tracing.

4. Assessment of automatic intonation stylization

4.1. Testing the model through resynthesis

If the perceptual model presented above is valid, it will preserve all the perceived prosodic information contained in the signal. It should therefore be possible to reconstruct a synthetic F0 contour which should prove indistinguishable from the original contour. In other words, if the listener is unable to distinguish the original F0 contour from that which is reconstructed from the stylized contour, this shows that no
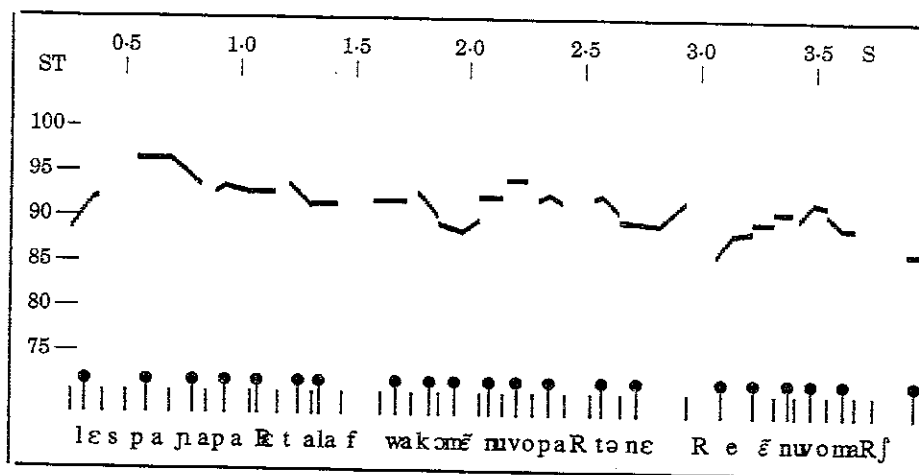
Figure 4. Tonal score for the sentence: "L'Espagne apparaît à la fois comme un nouveau partenaire, et un nouveau marché". Female speaker.

information is lost in the stylization process. The evaluation process is based on the measured F0 contours. It is the reason why F0 measurements and voicing decisions were carefully checked for the test stimuli. Therefore, we can assume that the validity of the assessment test was not affected by these types of errors.

The overall procedure to test this hypothesis then looks as follows. (1) Process the original F0 contour according to the model; this yields a stylized pitch contour. (2) Reconstruct an F0 contour starting from this stylized contour. (3) Resynthesize both the original and the modified F0 contours. (4) Compare both versions: they should be identical with respect to prosody.

F0 reconstruction is required because the stylization represents the perceived intonation, not F0 which is actually needed for resynthesis. The WTAP contour represents the result of the perceptual integration of an F0 contour; it is a perceptual object, and not a physical F0 contour. The use of this contour for resynthesis would imply a double application of perceptual integration. Therefore, it is necessary to remove the effect of perceptual integration prior to resynthesis. In other words, inversion of the stylized WTAP produces a physical F0 contour which in turn is supposed to be perceived in the same way as the stylized WTAP contour. The reconstruction process takes as input the stylized pitch contour, which is visualized in the tonal score. It delivers as its output a synthetic F0 contour. This synthetic contour is fully isomorphic to the tonal score. It can be shown that the inverse of Equation (1) is given by:

$$f(t) = p'(t) \int_0^t e^{-\alpha(t-\tau)} \, d\tau + p(t) \tag{5}$$

As a result one can compute a synthetic F0 contour, by application of Equation (5) to the stylized pitch contours, for each tonal segment.
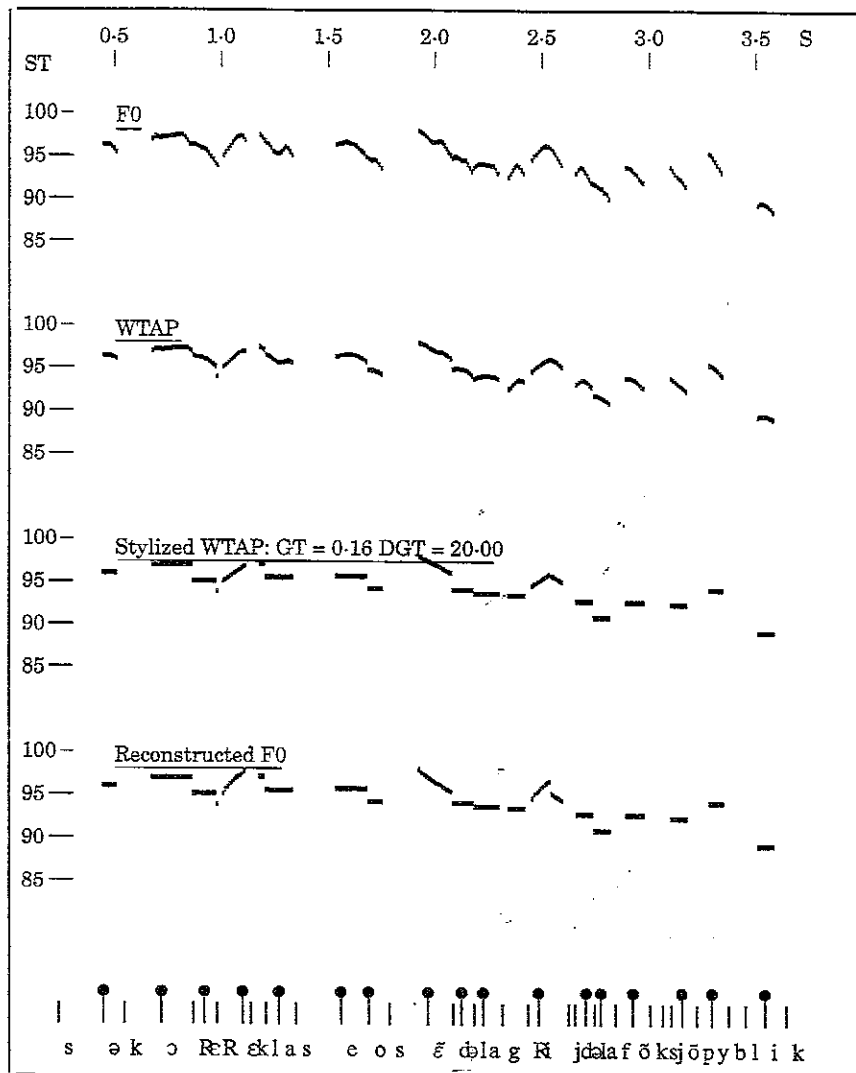
Figure 5. Automatic stylization. From top to bottom: F0contour, WTAP contour, stylized WTAP contour, reconstructed F0. Sentence: "Ce corps est reclassé au sein de la grille de la fonction publique". Female speaker.

Figs 5 and 6 give examples of the whole intonation stylization process. Four contours are presented, above the phonetic transcription and the phonetic segmentation and syllabification of the utterances. The first contour (top) is the F0 contour, the second contour is the WTAP contour. One can notice the smoothing effect of the WTAM. The third contour is the stylized WTAP contour. The values used for the glissando threshold (GT) and the differential glissando threshold (DGT) are indicated. The last contour is the reconstructed, stylized F0 contour, computed by application of Equation (5) to the
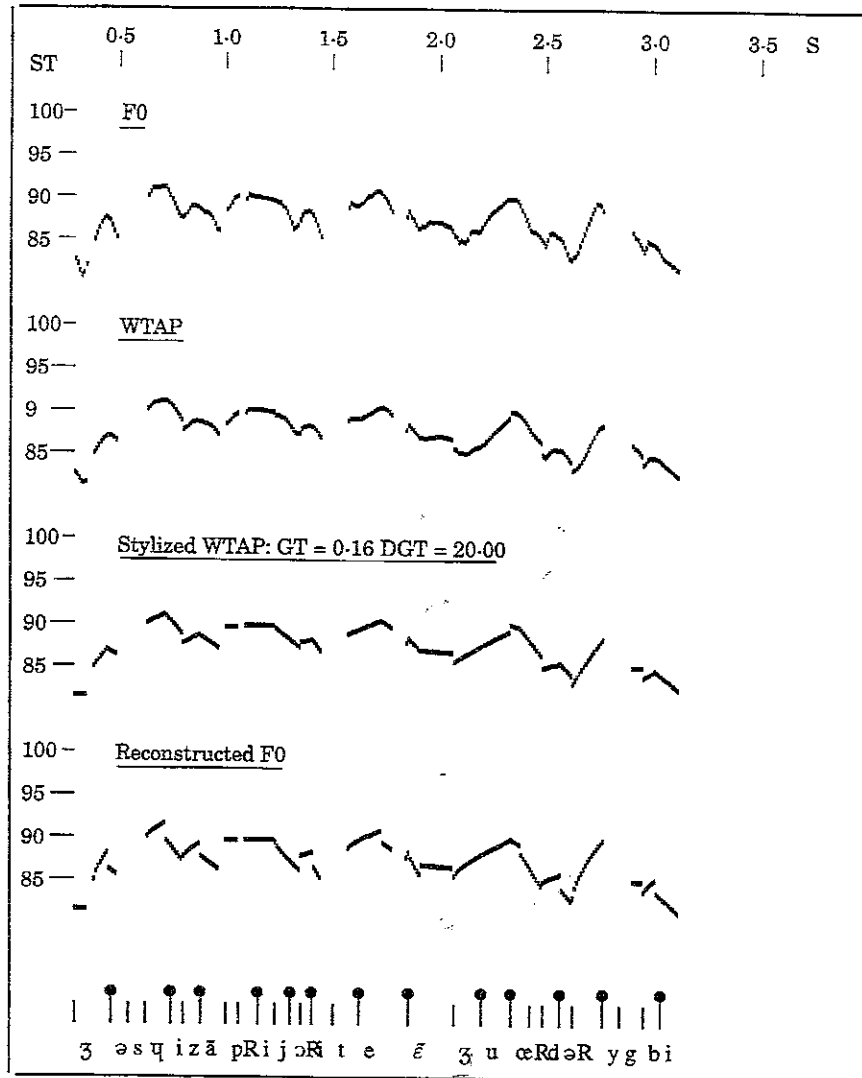
Figure 6. Automatic stylization. From top to bottom: F0contour, WTAP contour, stylized WTAP contour, reconstructed F0. Sentence: "Je suis en priorité un joueur de rugby". Male speaker.

stylized WTAP contour. The first and last contours are the natural and the stylized F0 contours, that were compared in the perception experiments.

Resynthesis was performed using a Time Domain Pitch Synchronous Overlap Add (TD-PSOLA for short) analysis/synthesis system (Mouline & Charpentier, 1989). Pitch periods were automatically marked, using a modified version of the algorithm presented by Dologlou and Carayannis (1989).

Although the quality of TD-PSOLA speech is comparable to that of natural speech, there is a slight difference between TD-PSOLA speech and natural speech, particularly

in presence of some additional background noise. We noticed that it was always possible to distinguish between TD-PSOLA processed signals and unprocessed signals, merely on the basis of subtle sound quality modifications. This is a potential problem for the test paradigm we used for stylization assessment. To avoid this kind of artefact, the original signals were also processed using TD-PSOLA (retaining the original pitch contour). The effect of this processing was to synthesize a new signal by adding pitch-synchronous overlapped windowed frames of the original signal. This produced a slight sound quality modification (because the sum of all the windows is not exactly one), but no pitch modification. As a result the speech signals under comparison have the same sound quality.

A preliminary test was run to check the quality obtained with the stylization algorithm and to determine the parameter settings for the final perception experiment.

In this pilot investigation, the glissando threshold was set according to Equation (2), and the differential glissando threshold was set to 20 (the values reported in the literature are between 12 and 40). The preliminary tests used informal listening to TD-PSOLA resynthesized utterances. The results seemed to indicate that almost all the stylized utterances were indistinguishable from the original utterances.

4.2. Speech material

The speech material for the pilot investigation and the experiment has been extracted from the following sources.

The first source was a large-scale database of read French speech, the BREF database. The text material was extracted from the newspaper "Le Monde". A description of the design and content of the BREF database can be found in Gauvain, Lamel and Eskénazi (1990), Lamel, Gauvain and Eskénazi (1991). Speech signals with reliable phonetic segmentations were available for 120 speakers. Utterances were selected randomly from this database.

The second source is a set of 60 sentences read by one male speaker, for which a manual close-copy (straight-line) stylization and a manual phonetic segmentation are available (Beaugendre, d'Alessandro, Lacheret-Dujour & Terken, 1992). This corpus was originally designed for an intonation study in the context of Text-to-Speech synthesis.

For all signals, the sampling rate was 16 kHz, with a 16 bit resolution. The speech material has been digitally recorded directly on the computer mass storage device.

4.3. Method

The aim of this formal testing was to measure the perceptual proximity between TD-PSOLA resynthesized speech with natural intonation contours and stylized contours. As preliminary experiments indicated that natural and stylized contours were probably indistinguishable, a same/different test paradigm seemed appropriate.

4.3.1. Stimuli

The test corpus consisted of 30 utterances, in two sets. The first set contained 25 utterances pronounced by 10 speakers, five males and five females. Both the speakers and the utterances were selected at random from part C of the BREF database. These

TABLE I. Stylization parameters used in the perception experiment

Stimulus	Glissando	Diff. glissando
V1	—	—
V2	0·16	20
V3	0·32	20
V4	0·64	60

utterances were relatively long sentences with an average duration of 4·575 s, and an average number of words, syllables and phonemes of respectively 12, 21·6 and 45·8. The second set were five relatively short utterances, with an average duration of 1·5 s and an average of five words pronounced by an additional male speaker. They were selected from the close-copy database mentioned above.

For each sentence in the test corpus, four TD-PSOLA resynthesized versions were computed. The first one (V1) had the original F0 contour. The other versions (V2, V3, V4) had stylized F0 contours, using different values for the glissando and the differential glissando thresholds, as indicated in Table I. For the glissando threshold, these values correspond to the numerator in Equation (2). For the differential glissando threshold, these values represent the difference of glissando rates $(g2-g1)$, as in Equation (4).

As can be seen from this table, thresholds increase with version numbers. In V2 the glissando threshold G_{tr} is as in Equation (2), which is the measured threshold for isolated pure tones. In V3, G_{tr} is doubled, as would be the case if the pitch change of an isolated pure tone would need to be twice as large in order to be judged as dynamic by the subjects. In V4, G_{tr} is doubled compared to V3, and the differential glissando threshold is multiplied by a factor of three.

As a result, the stylization will produce less dynamic tones with the settings of V3 and even less with those of V4. As a matter of fact, almost all tones in V4 will be static. Fig. 8 illustrates the WTAP contours V1, V2, V3 and V4 used in the experiments. Fig. 9 represents narrow-band spectrograms of a few syllables contrasting stimuli V1, V2, V3, and V4.

The parameter settings for V3 and V4 were chosen for the following reasons. First, following the argument of de Pijper (1983), it was necessary to introduce some clearly different stimuli among the versions V1, V2, V3 and V4, in order to avoid a bias in testing. Indeed, if all stimuli are very similar, subjects can be inclined to answer "different" at random just to avoid having to answer "same" in all cases.[5] Second, as the sets V2, V3 and V4 were chosen according to the perceptual thresholds for isolated pure tones, the results could provide an indirect verification of the relevance of these thresholds for continuous speech.

Starting from these four versions (V1, V2, V3 and V4) of each sentence, four categories of stimuli were constructed: V1V1, V1V2, V1V3 and V1V4. Each stimulus consisted of a pair of two versions of the same utterance, separated by a 750 ms silent interval. This resulted in a set of $(30 \times 4 =)$ 120 stimuli.

The subjects listened to each stimulus. They were asked to indicate whether the members of a stimulus pair were identical or not. In order to do this they had to choose

[5] As a matter of fact, we think that this type of random response was actually given by some subjects.
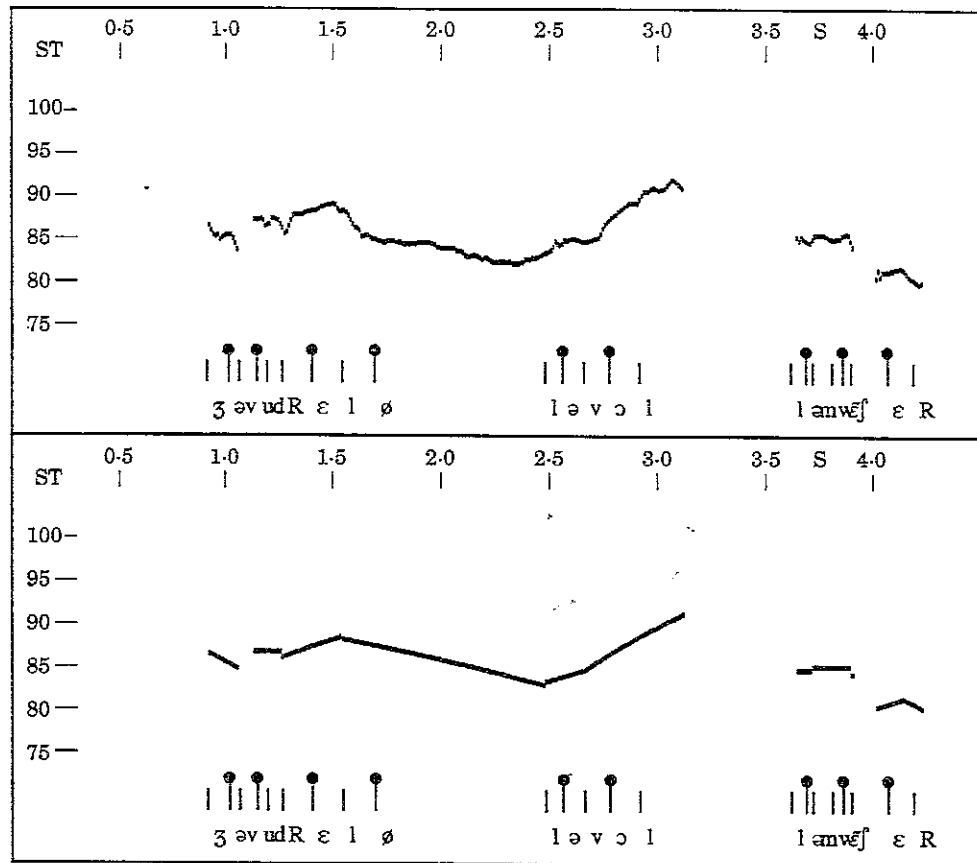
Figure 7. Top: F0contour and syllabic segmentation. Bottom: Tonal score. Spontaneous speech. Sentence: "Je voudrais le. .-. le vol, le moins cher". Male speaker.

the response "same" or "different". They were not told to pay special attention to intonation. The question asked to the subjects was expressed as follows:

> You will listen to pairs of sentences. Each pair contains either two identical sentences, or one natural and one modified sentence. The goal of this experiment is to determine if members of a pair are the same sentence or not. If you can hear a difference between the two sentences (in terms of intonation, segments, sound quality, etc.), please answer "NO". If the two sentences of a pair are strictly the same, please answer "YES".

4.3.2. Test procedure

The test procedure was computerized, using a software environment described in Lamel (1991). Stimuli were played in a random order, and members of a stimulus pair were also played in a random order (e.g. V1V2 or V2V1). The number of stimuli proposed
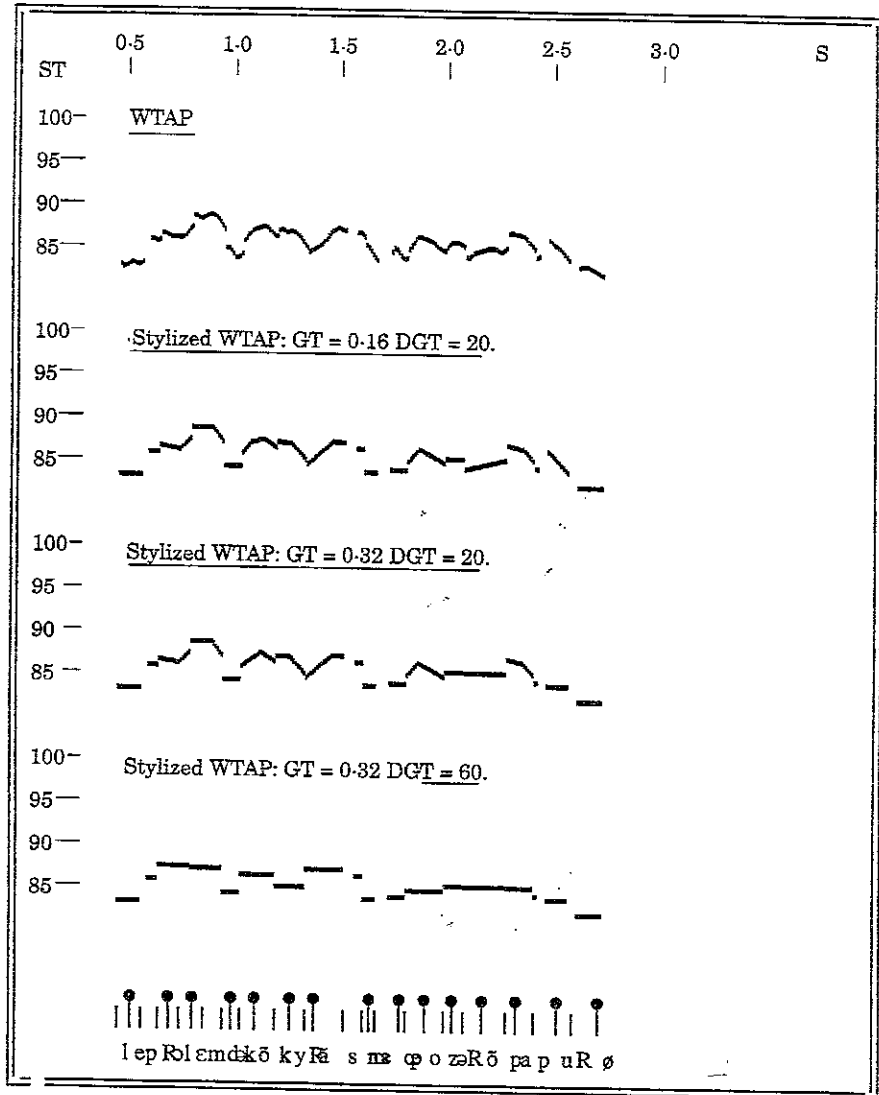
Figure 8. Comparison of stylized contours for the experiments. From top to bottom: natural contour (V1), first stylized contour (V2), second stylized contour (V3), third stylized contour (V4). Sentence: "Les problèmes de concurrence ne se poseront pas pour eux". Male speaker.

in one test session was also chosen at random, with an average of 52 stimuli pairs.

The first five stimuli at the beginning of each experimental session were training stimuli, and the answers for these stimuli were not included in the statistics. Therefore, an average of 47 responses were actually used for each session. Four sessions were assigned to each subject, who listened to an average of 208 stimuli pairs (an average of 188 stimuli pairs were actually used for each subject).

A total number of 3776 (20 subjects, and an average of 188 responses per subject)
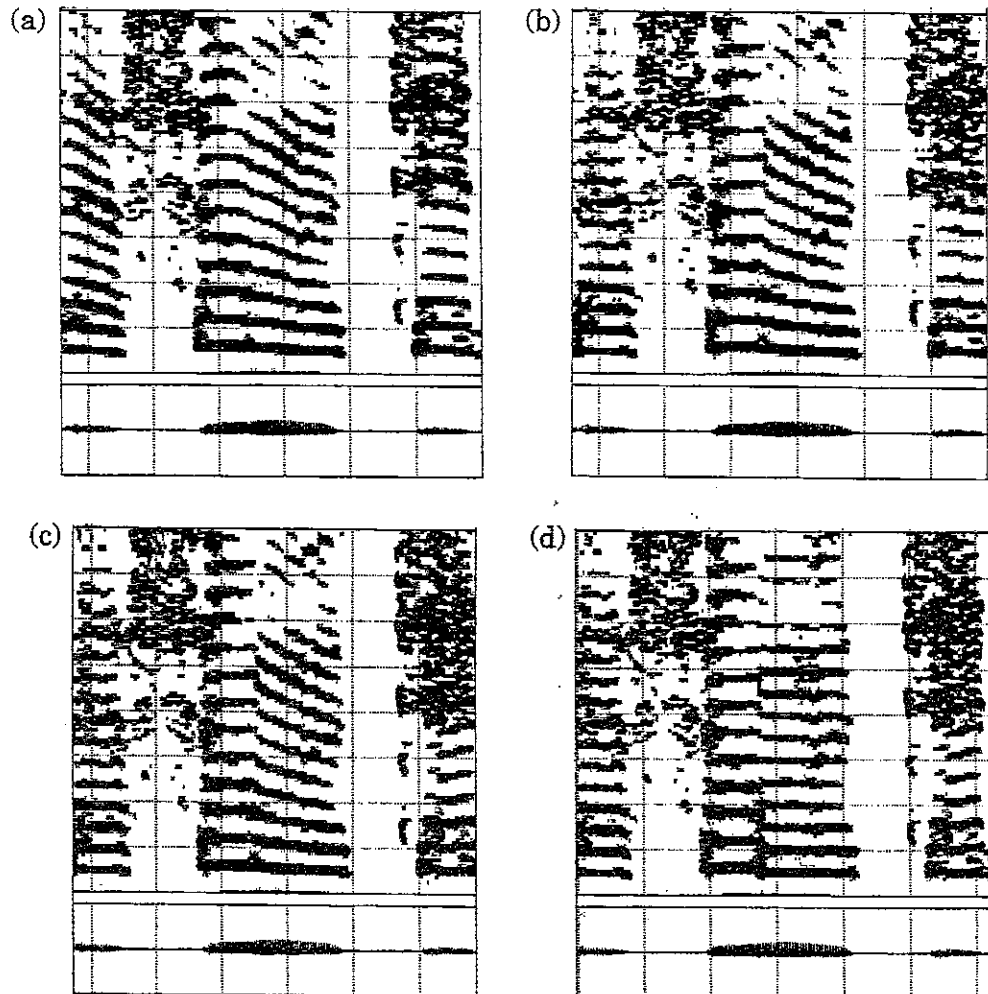
Figure 9. Comparison of narrow-band spectrograms of a few syllables. A: Natural contour (V1); B: first stylized contour (V2); C: second stylized contour (V3); D: third stylized contour (V4). Sentence: "La fermeté". Female speaker.

responses were available for computing statistics. Each session lasted about 10–15 min, and the four test sessions lasted about 1 h. The four test sessions were typically spread over 1–4 weeks.

The subjects answered by pointing at colored boxes on the computer screen with a mouse device. The sound stimuli were presented monaurally through Beyer Dynamic DT 48 headphone, directly from the computer memory, at 80 dB SPL.

The 20 subjects were members of the laboratory. Some of them already participated in (several) psychoacoustic and speech perception experiments. Some of the subjects can also be considered as experts in phonetics or automatic speech processing. All subjects were native speakers of French, and had no known hearing loss. Tone

TABLE II. Results for the same/different experiment. Percentage of correct responses and number of settings for categories V1V1, V1V2, V1V3, V1V4

Subject	% Correct V1V1	% Correct V1V2	% Correct V1V3	% Correct V1V4
1. AB	89·13 (46)	46·51 (43)	62·00 (50)	87·24 (47)
2. CDA	90·70 (43)	27·69 (65)	43·90 (41)	96·43 (28)
3. CN	90·70 (43)	26·09 (46)	32·65 (49)	66·67 (42)
4. EG	96·08 (51)	7·02 (57)	22·73 (44)	48·84 (43)
5. GA	100·00 (47)	30·00 (50)	36·00 (50)	69·57 (46)
6. JFK	100·00 (44)	5·27 (57)	20·00 (45)	38·46 (39)
7. JSL	69·23 (52)	42·86 (49)	26·00 (50)	76·19 (42)
8. LD	92·00 (50)	43·40 (53)	50·00 (42)	84·09 (44)
9. MGR	94·64 (56)	47·82 (46)	82·69 (52)	93·88 (49)
10. OV	86·11 (36)	29·83 (57)	26·00 (50)	38·46 (39)
11. PB	85·46 (55)	50·94 (53)	61·70 (47)	83·33 (42)
12. PBM	90·25 (41)	38·46 (52)	63·64 (44)	85·71 (49)
13. SR	71·43 (35)	32·79 (61)	40·00 (45)	51·22 (41)
14. SG	95·83 (48)	1·92 (52)	9·62 (52)	10·87 (46)
15. SF	92·86 (56)	32·14 (56)	32·50 (40)	90·91 (44)
16. VP	86·67 (45)	28·57 (56)	15·55 (45)	40·48 (42)
17. BD	84·44 (45)	36·96 (46)	54·90 (51)	81·25 (48)
18. MJ	93·48 (46)	23·40 (47)	27·91 (43)	42·31 (52)
19. MAD	85·11 (47)	43·75 (48)	62·75 (51)	85·00 (40)
20. FB	97·62 (42)	59·32 (59)	85·72 (42)	94·87 (39)
All subjects	89·76 (928)	32·38 (1053)	42·87 (933)	67·98 (862)

audiograms were performed after completion of the test, for 15 subjects out of 20:[6] they have all normal hearing (tone thresholds in the range −10+30 dB ISO for all the 15 subjects, for frequencies between 250 and 8000 Hz).

4.4. Results and discussion

First of all, the responses can be analysed in terms of their correctness. Since the aim of the test was to see whether subjects could determine if a pair of utterances were identical or not, and since only the pairs in category V1V1 were identical, the correct response was "same" for category V1V1, and "different" for categories V1V2, V1V3, V1V4.

The percentages of correct responses are shown in Table II organized by stimulus category (V1V1, V1V2, V1V3, V1V4) and by subject. The row "All Subjects" shows the average of the results of all subjects. For each subject and each category, the table shows the percentage of correct responses, along with the total number of responses between parentheses.

Two types of results were expected. On the one hand, a high score of correct responses was expected in categories V1V1 and V1V4, as the signals pairs in these categories were respectively identical and clearly different. These "clear cases" were included in the test material in order to check the ability of the subject to do the task. On the

[6] Subjects 1, 2, 3, 4, 5, 7, 8, 9, 11, 12, 13, 16, 17, 18, 19. The five other subjects left the laboratory after completion of the test.

TABLE III. Statistical analysis of the results for the same/different experiment

Subject	Z-score H1	Z-score H2	χ^2 H3	Z-score H4	Z-score H5
1. AB	−0·20 A	1·24 A	56·39 R	−0·87 A	−3·26 R
2. CDA	0·15 A	2·17 A	26·05 R	2·03 A	−0·50 A
3. CN	0·15 A	−2·16 R	15·36 R	1·93 A	−0·17 A
4. EG	1·45 A	−5·11 R	1·45 R	5·08 A	3·14 A
5. GA	2·29 A	−1·77 R	424·76 R	1·44 A	−0·82 A
6. JFK	2·21 A	−6·49 R	10·48 R	5·35 A	3·44 A
7. JSL	−4·99 R	−0·62 A	3·36 A	−0·41 A	−2·89 R
8. LD	0·47 A	0·68 A	90·25 R	−0·50 A	−3·09 R
9. MGR	1·16 A	2·43 A	163·62 R	−1·08 A	−3·57 R
10. OV	−0·78 A	−6·49 R	12·11 R	1·57 A	−0·84 A
11. PB	−1·12 A	0·54 A	56·50 R	−1·63 A	−4·36 R
12. PBM	0·05 A	1·00 A	48·67 R	0·23 A	−2·24 R
13. SR	−3·66 R	−4·61 R	0·53 A	1·15 A	−1·40 A
14. SG	1·35 A	−11·72 R	0·65 R	5·60 A	3·84 A
15. SF	0·71 A	1·81 A	52·76 R	1·20 A	−1·23 A
16. VP	−0·74 A	−6·40 R	11·26 A	1·75 A	−0·62 A
17. BD	−1·24 A	0·22 A	16·04 R	0·42 A	−1·87 R
18. MJ	0·79 A	−6·80 R	21·98 R	2·32 A	0·25 A
19. MAD	−1·12 A	0·79 A	31·54 R	−0·53 A	−3·00 R
20. FB	1·65 A	2·32 A	823·05 R	−3·03 R	−6·09 R

"R" means that $P<0·01$ (the results differ highly significantly from the hypothesis), "r" means that $P<0·05$ (the results differ significantly from the hypothesis), "A" means that $P>0·05$ (the results do not differ significantly from the hypothesis). See text for explanation of the hypotheses H1–H5.

other hand, a lower score (or a higher percentage of errors) was anticipated for categories V1V2 and V1V3, where some confusion between natural and stylized contours was to be expected. The difference between the scores in V1V2 and V1V3 will enable us to evaluate the similarity between stylized and natural intonation contours.

The first aspect, the ability of the subjects to do the discrimination task, can be evaluated by inspecting the responses for categories V1V1 and V1V4. Ideally, all stimuli in category V1V1 should be judged identically, and most of the stimuli in category V1V4 should be judged differently. The average score (all subjects pooled) of 89·76% correct responses in category V1V1 indicates a general trend for these stimuli to be judged identically. Nevertheless, some individual variation is noticeable, as the score is varying between 100 and 69·23%. The relatively high scores in category V1V4 indicate that the differences between the two signals in these pairs were not large enough in order to be distinguished by all subjects. Still, some subjects were able to distinguish these signals (on average 67·98% of correct responses). Again, some individual variation is noticeable, as the score is varying between 10·87 and 96·43%. The variation is much larger within category V1V4, compared to category V1V1. It was not possible to generate stimuli that would have been more different, only by changing the perceptual thresholds for the stylization: almost all tones in V4 stimuli were static tones. It would have been possible to introduce artificial degradation for this stimuli, but we preferred to generate all stimuli using our stylization algorithm.

Table III shows some data on the statistical significance of the results. For category

V1V1, we tested the null hypothesis H1 that more than 90% of the responses are correct. Z-scores for a unilateral test are reported in column 2 of the table. This assumption proved to be statistically true for a large majority of subjects. The critical Z-scores are -2.33 and -1.645 at significance levels 0.01 and 0.05 respectively ("R" and "r" indicate that H1 can be rejected at significance levels 0.01 and 0.05 respectively. "A" indicates that H1 can be accepted). This is a first indication of the ability of the subjects to perform the task. For category V1V4, we tested the null hypothesis H2 that more than 80% of the responses are correct. Again, Z-scores for a unilateral test are shown in Table III, and this assumption proved to be statistically true for a majority of subjects. This is a second indication on the subjects performance. We can therefore conclude that most of the subjects were successful at doing the task.

For category V1V2, the aim of the test was to check the perceptual equality of natural and stylized contours. A high percentage of confusion (incorrect responses) was expected. This percentage is indeed high. The null hypothesis H3 that the number of correct responses for category V1V1 and the number of incorrect responses for V1V2 are identical was tested using a χ^2 test. This hypothesis must be rejected for a majority of subjects, although it can be accepted for some subjects (see Table III, "R" and "r" indicate that the hypothesis is rejected at the significance levels 0.01 and 0.05 respectively. "A" indicates that the hypothesis cannot be rejected). This indicates that the majority of subjects were able to distinguish the sentences in the V1V2 pairs better than in the V1V1 pairs.

This does not mean that the stylization procedure is not efficient. The hypothesis H4 that more than 60% of the stylized and natural contours are confused is statistically true for all the subjects, except one. The hypothesis H5 that more than 75% of the stylized and natural contours are confused is still statistically true for a majority of the subjects. It should be stressed once again that subjects were not asked to detect differences in intonation, but rather differences of any kind between signals. Many subjects reported that they could distinguish signals on the basis of changes in some aspects of sound quality rather than on the basis of differences in intonation. As for category V1V3, the situation is in between category V1V2 and V1V4, as expected (see Tables II and IV).

Table IV shows the same data as Table II for the five shorter sentences. The results are along the same lines, except that more confusions occurred in category V1V2. On average (All Subjects) 91.49% of correct responses in category V1V1, and 21.08% of correct responses (78.92% of confusion) in category V1V2. Nevertheless, there are not enough settings available for computing statistics for the short sentences.

The responses of the subjects, which were analysed in terms of their correctness, can also be plotted as a function of the number of answers "SAME", as in Table V. In this case, the proportion of "SAME" answers gradually decreases from condition V1V1 to condition V1V4, showing that the subjective similarity decreases as the thresholds (i.e. the model parameters) increase.

A by product of this experiment is the indirect evaluation of the importance of perceptual thresholds. It must be emphasized that even with a very crude stylization, such as in category V1V4, some subjects made a lot of confusions. This might indicate that decomposition of intonation contours into short-duration tones provides a rather good basis for robust intonation stylization. Moreover, the percentage of correct responses clearly correlates with the magnitude of the perceptual thresholds used. The thresholds are therefore meaningful, but probably not very critical.

TABLE IV. Results for the same/different experiment

Subject	% Correct V1V1	% Correct V1V2	% Correct V1V3	% Correct V1V4
1. AB	80·00 (5)	75·00 (4)	55·56 (9)	85·71 (7)
2. CDA	100·00 (6)	20·00 (10)	42·86 (7)	100·00 (9)
3. CN	100·00 (7)	0·00 (5)	7·14 (14)	66·67 (3)
4. EG	100·00 (10)	0·00 (11)	0·00 (5)	66·67 (3)
5. GA	100·00 (8)	11·11 (9)	16·67 (6)	100·00 (6)
6. JFK	100·00 (5)	0·00 (10)	0·00 (7)	75·00 (4)
7. JSL	85·71 (7)	14·29 (7)	16·67 (12)	100·00 (7)
8. LD	100·00 (5)	16·67 (6)	33·33 (9)	100·00 (7)
9. MGR	100·00 (7)	14·29 (7)	71·43 (7)	75·00 (8)
10. OV	80·00 (5)	28·57 (7)	0·00 (7)	36·36 (11)
11. PB	87·50 (8)	71·43 (7)	100·00 (9)	80·00 (10)
12. PBM	50·00 (4)	57·14 (7)	71·43 (7)	75·00 (4)
13. SR	66·67 (3)	0·00 (9)	22·22 (9)	0·00 (5)
14. SG	100·00 (9)	0·00 (14)	0·00 (8)	12·50 (8)
15. SF	76·92 (13)	25·00 (4)	41·67 (12)	83·33 (12)
16. VP	100·00 (7)	45·45 (11)	0·00 (8)	33·33 (3)
17. BD	80·00 (5)	27·27 (11)	78·57 (14)	75·00 (4)
18. MJ	100·00 (9)	11·11 (9)	11·11 (9)	50·00 (8)
19. MAD	92·31 (13)	22·22 (9)	66·67 (6)	100·00 (4)
20. FB	100·00 (5)	33·33 (9)	50·00 (6)	66·67 (3)
All subjects	91·49 (141)	21·08 (166)	35·09 (171)	69·84 (126)

Short sentences. Percentage of correct responses and number of settings for categories V1V1, V1V2, V1V3, V1V4.

TABLE V. Average number of "SAME" answers as a function of stimulus type and stimulus

Stimulus type	V1V1	V1V2	V1V3	V1V4
Long	89·76	67·62	57·13	32·02
Short	91·49	78·92	64·91	30·16

House (1990, p. 115), who used automatic straight line stylization, reports that:

> the majority of the stylized sentences could not be distinguished from their original counterparts on the basis of intonation alone.

It seems very difficult to design a test paradigm that is able to measure discrimination of sentences on the basis of intonation alone. It is for this reason that the same/different paradigm was preferred, even if it is much more severe. Our results indicate that, using the stylization procedure described above, a (large) majority of sentences could not be distinguished at all.

It is of interest to compare the results for hand-made straight line (HMSL, for short) stylization experiments reported in the literature (e.g. de Pijper, 1983; t'Hart *et al.*, 1990; Beaugendre *et al.*, 1992), with those obtained with automatic tonal stylization. Extrapolating the data (correct responses) from de Pijper, we obtain 95·5% in category

V1V1 (natural/natural) and 13·2% in category V1V2 (natural/best stylization), for all subjects results, to be compared to 89·76 and 32·38% for all sentences, 91·49 and 21·08% for short sentences, also for all subjects. For short sentences, the scores are almost comparable. As a matter of fact, de Pijper used short sentences (2–3 s), the duration of which was somewhere in between that of the short and long sentences we used. The differences between the results of these two experiments can be due to differences in the experimental conditions. First, two different types of signal processing (TP-PSOLA vs. LPC) are used. In de Pijper's study, some small signal differences may have been masked by the poor quality of LPC speech. Second, the question asked to the subjects was also different. In our case subjects were asked to discriminate two signals, without restrictions on the nature of the differences. In the HMSL stylization tests, subjects were always asked to concentrate on intonation, and not on other aspects of the signal (e.g. de Pijper, 1983, pp. 117–118). Another difference is that our test material contained both male and female voices, whereas the HMSL stylization experiments used male voices only.

On the basis of the outcome of this experiment, we feel entitled to conclude that it is possible to automatically stylize French intonation, while maintaining a high level of perceptual equality. Moreover, the stylization procedure turns out to be quite robust, since many of the V1V3 stimuli were judged identical, indicating that in many cases, the subjects were unable to spot the impact of the stylization. In our opinion, this robustness is mainly due to the segmentation step: when intonation contours are divided into short tones according to the segmental information, and when the pitch perceived for these short tones is computed on the basis of an average of the measured fundamental frequency values, in many cases errors in static/dynamic tones decision are only of secondary importance.

This does not mean that these errors are not important. The robustness of the decomposition might vary as a function of the language under study.

5. Discussion and conclusion

5.1. Summary

In this study we presented a computer model of tonal perception in speech, and we described a perception experiment aiming at evaluating this model. The algorithm for automatic stylization of pitch contours was actually used to generate the test material. The model and the algorithm require five major processing steps.

(1) In the parametric analysis, all relevant acoustic parameters of the speech signal are determined: fundamental frequency, voiced/unvoiced decision.

(2) Next, a segmentation of the speech signal into phonetic syllables is obtained, on the basis of phonetic labels supplied by a speech recognition system. The segmentation outputs the sequence of voiced speech portions corresponding to individual phonetic syllables in the speech signal under analysis.

(3) At this point, we can go from the acoustic to the perceptual domain. The first step is to simulate the short-term pitch integration, by calculating the weighted time–average pitch which will be used later for the stylization.

(4) Each syllable pitch contour is subdivided into one or more tonal segments. For this purpose, the contour is decomposed into uniform parts. Perceptual thresholds

are used to decide when to stop the segmentation, and which parts to merge into single tonal segments.

(5) Finally the pitch contour is stylized by selecting the WTAP of the start and end points of each tonal segment, and by linearly interpolating between them.

Some properties of the model will be discussed below.

5.2. Properties of the model

The model proposed above has some interesting properties.

Most importantly, it is a perceptual model. Whereas in earlier systems ad hoc heuristics are used for many aspects, here perceptual processing is simulated as much as possible. As the stylization is controlled by two parameters, which are perceptual thresholds, it can be used to measure these thresholds (at least for signals with an obvious syllabic segmentation). This gives the model a scientific interest: it becomes a tool for basic research on tonal perception. Any model can be verified in perceptual experiments. But unlike other systems, which can only claim that the obtained stylization is perceptually equivalent to the natural pitch contour (and hence descriptively adequate), our system also explains why this is so (it also has explanatory adequacy).

Because the stylization is fully automatic, there is no bias whatsoever from the user phonetician. In this respect the system has a clear advantage over the close-copy stylization procedure.

Another asset is the clear distinction between acoustic, perceptual, and linguistic representations of intonation. The importance of this distinction has been stressed in the introduction and in the discussion on the concepts of tones, tonal segments, and pitch targets. The stylized contour reflects the signal after low level perceptual processing, prior to any categorization involving a language-specific intonation grammar. As a result, this auditory representation can be defined and investigated (i.e. measured) on its own, without reference to some abstract intonation model, or even without reference to the communicative function of pitch in speech.

A key feature of our method is that very little assumptions are made on the linguistic aspects related to intonation. The latter are known to be language-dependent. As a result, this method could serve as a basis for future work on language-independent intonation analysis.

To conclude, the model presented above is a first attempt towards a complete model of intonation perception. However, there are still some intrinsic limitations with this model.

A fundamental weakness comes from our current limited knowledge on the perception of prosody. The currently available data on the perception of prosody is still too fragmentary to justify each of the steps in the model. This can be seen in two areas.

Firstly, the model relies on the preliminary decomposition of the pitch contour into short duration tones. While there are indications that such a segmentation does indeed take place, no quantitative data are available on the impact of spectral changes, and energy changes on the perception of pitch (see below: future work). We relied on the crude hypothesis that phonetic change can produce intonation segmentation. This hypothesis is probably justified for abrupt changes (e.g. stops, changes in voicing, pauses). But for smooth transitions (e.g. involving liquids, nasals, semi-vowels, or contiguous vowels) the segmentation is less obvious. The method relies on the phonetic

labelling of the utterance in order to determine (the voiced parts of) the phonetic syllables. However, the way in which this phonetic segmentation is to be achieved is not described in the paper. Of course it requires a fully-fledged speech recognition system to do this (at least in those applications where phonetic alignment is not possible), but this lies outside the scope of this study. Moreover, automatic phonetic alignment is now a fairly common tool in speech research. Probably a less sophisticated segmentation algorithm could do (e.g. Mertens (1987*b*)).

Secondly, the model assumes a differential slope threshold, for which the data in the literature is extremely scarce. As a result, the threshold value may seem to be chosen arbitrarily. However, the selected value was found to be useful for the experiments.

The model was tested for one language only, i.e. contemporary French as spoken in France; and this could be seen as a limitation. It is clear that the segmental and suprasegmental properties of French may favour a certain approach which could be less successful for other languages. For instance, syllabic decomposition is an important feature of French (when compared to English, say), and the set of possible pitch movements is rather limited (again when compared to English).

Finally, the model does not include the last processing step mentioned in Section 2. As a result, micro-prosodic variation is not normalized, and stress is not determined. The model provides no linguistic interpretation of the data, because it would require language-specific information.

5.3. Future work

In order to improve the accuracy of the model, additional perceptual experiments should be run to determine the thresholds related to tonal phenomena in continuous speech. More particularly the glissando and differential glissando thresholds for continuous speech should be measured, to see to what extent they differ from the thresholds for isolated stimuli without change in spectral envelope.

Another set of experiments should be undertaken to investigate the perceptual segmentation of the speech signal due to spectral change and loudness variation. A possible experiment to investigate the role of amplitude variation would be to compare the perceived glissandi for sounds with and without intensity variation. The aim of the experiment could be to answer the following question. For a sound with changing pitch and decreasing or increasing intensity, which part of the variation will be perceived, depending on the amount of intensity change? In order to answer this question, one could use a same/different test on a pair of stimuli with identical pitch change, where the first stimulus has a constant intensity whereas for the other stimulus intensity decreases or increases. This would enable one to determine the minimal amplitude change required to affect tonal perception.

The model for tonal perception has several immediate applications, some of which have already been implemented for the described experiment. The first is fully automatic analysis of intonation, with representation of the perceived tonal score. The second is the specification of prosody for speech synthesis. Another future application is the automatic transcription of intonation, of which the tonal perception model would be a first but major step.

their kind help in the course of this research. The authors are also grateful to the two anonymous reviewers for their useful comments.

References

d'Alessandro, C. & Castellengo, M. (1994). The pitch of short-duration vibrato tones. *Journal of the Acoustical Society of America* 95, 1617–1630.

Bagshaw, P. C. (1993). An investigation to acoustic events related to sentential stress and pitch accents, in English. *Speech Communication* 13, 333–342.

Beaugendre, F., d'Alessandro, C., Lacheret-Dujour, A. & Terken, J. (1992). A perceptual study of French intonation. *Proceedings of International Conference on Speech and Language Processing*, Banff, 379–382.

Bosch, L. ten (1993). Algorithmic classification of pitch movement. *Proceedings of an ESCA Workshop on Prosody*, Working Papers 41, Lund University, Dept. of Linguistics, Lund, 242–245.

Carbonell, N. & Laprie, Y. (1993). Automatic detection of prosodic cues for segmenting continuous speech into supralexical units. *Proceedings of an ESCA Workshop on Prosody*, Working Papers 41, Lund University, Dept. of Linguistics, Lund, 184–187.

Dologlou, I. & Carayannis, G. (1989). Pitch detection based on zero-phase filtering. *Speech Communication* 8, 309–318.

Gauvain, J. L., Lamel, L. F. & Eskénazi, M. (1990). Design considerations and text selection for BREF, a large French read-speech corpus. *Proceedings of International Conference on Speech and Language Processing*, Kobe, 24.6.1–4.

Geoffrois, E. (1993). A pitch contour analysis guided by prosodic event detection. *Proceedings of Eurospeech 93*, Berlin.

Gibbon, D. & Braun, G. (1988). The PSI/PHI architecture for prosodic parsing. *Proceedings of the 12th International Conference on Computational Linguistics (COLING 88)*, Budapest, 202–204.

t'Hart, J. (1976). Psychoacoustic backgrounds of pitch contour stylization. *IPO—Annual Progress Report 11*, Eindhoven, 11–19.

t'Hart, J. (1979). Explorations in automatic stylization of F0 curves. *IPO—Annual Progress Report 14*, Eindhoven, 61–65.

t'Hart, J., Collier, R. & Cohen, A. (1990). *A Perceptual Study of Intonation*, Cambridge University Press, Cambridge.

t'Hart, J. (1991). F0 stylization in speech: straight lines versus parabolas. *Journal of the Acoustical Society of America* 90, 3368–3370.

Hermes, D. J. (1988). Pitch measurement by subharmonic summation. *Journal of the Acoustical Society of America* 83, 257–264.

Hess, W. (1983). *Pitch Determination of Speech Signals. Algorithms and Devices*. Springer Verlag, Berlin.

Hirst, D. J., Nicolas, P. & Espesser, R. (1991). Coding the F0 of a continuous text in French: an experimental approach. *Proceedings of the International Congress of Phonetic Sciences*, Aix-en-Provence, 234–237.

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

Huber, D. (1990). Prosodic transfer in spoken language interpretation. *Proceedings of the International Conference on Speech and Language Processing*, Kobe, Japan, 1, 12.7.1–4.

Klatt, D. H. (1973). Discrimination of fundamental frequency contours in synthetic speech: implications for models of pitch perception. *Journal of the Acoustical Society of America* 53, 8–16.

Kloker, D. R. (1976). A technique for the automatic location and description of pitch contours. *Proceedings of the IEEE Symposium on Acoustics, Speech, and Signal Processing*, Philadelphia, Pennsylvania.

Kohler, K. J. (1991). Prosody in speech synthesis: the interplay between basic research and TTS application. *Journal of Phonetics* 19, 121–138.

Lamel, L. (1991). FCtest software and associated ScoreFC scoring software. *LIMSI report NDL 91–22*.

Lamel, L. F. & Gauvain, J. L. (1993). High performance speaker-independent phone recognition using CDHMM. *Proceedings Eurospeech 93*, Berlin.

Lamel, L. F., Gauvain, J. L. & Eskénazi, M. (1991). BREF, a large vocabulary spoken corpus for French. *Proceedings of EuroSpeech 91*, Genova, 505–508.

Lea, W. A., Medress, M. F. & Skinner, T. E. (1975). A prosodically guided speech understanding system. *IEEE Transaction on Acoustics, Speech and Signal Processing* 23, 30–38.

Martin, P. (1982). Comparison of pitch detection by cepstrum and spectral comb analysis. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 180–183.

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

Mertens, P. (1987b). Automatic segmentation of speech into syllables. *Proceedings of the European Conference on Speech Technology*, Laver & Jack, Eds. Edinburgh: CEP Consultants, Vol. II, 9–12.

Mertens, P. (1989). Automatic recognition of intonation in French and Dutch. *Proceedings of Eurospeech 89*, Paris 1, 46–50.

Mouline, E. & Charpentier, F. (1989). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication 9*, 453–468.

Nishinuma, Y. (1979). Un modele d'analyse automatique de la prosodie. Accent et intonation en japonais. *Collection 'Sons et Parole'*, 1, Editions du CNRS, Paris.

de Pijper, J. (1983). *Modelling British English Intonation*. Cinnaminson Foris, Dordrecht.

Pollack, I. (1968). Detection of rate of change of auditory frequency. *Journal of Experimental Psychology 77*, 535–541.

Rietveld, A. C. M. (1984). Syllaben, klemtonen en de automatische detectie van beklemtoonde syllaben in het Nederlands. Unpublished doctoral dissertation, University of Nijmegen.

Rossi, M. (1971). Le seuil de glissando ou seuil de perception des variations tonales pour les sons de la parole. *Phonetica 23*, 1–33.

Rossi, M. (1978). La perception des glissando descendants dans les contours prosodiques. *Phonetica 35*, 11–40.

Rossi, M., Di Cristo, A., Hirst, D., Martin, P. & Nishinuma, Y. (1981). *L'intonation. De l'acoustique à la sémantique*. Klincksieck, Paris.

Schouten, H. E. M. (1985). Identification and discrimination of sweep tones. *Perception and Psychophysics 37*, 369–376.

Seashore, C. E. (1938). *Psychology of Music*. Dover, New York.

Sergeant, R. L. & Harris, J. D. (1962). Sensitivity to unidirectional frequency modulation. *Journal of the Acoustical Society of America 34*, 1625–1628.

Vaissière, J. (1988). The use of prosodic parameters in automatic speech recognition. *Recent Advances in Speech Understanding and Dialog Systems*. Niemann, Lang & Sagerer, Eds, pp. 71–99. Springer, Heidelberg.

Waibel, A. (1988). *Prosody and Speech Recognition*. Pitman, London.

Watson, C. S., Foyle, D. C. & Kidd, G. R. (1990). Limited processing capacity for auditory pattern discrimination. *Journal of the Acoustical Society of America 88*, 2631–2638.

Wightman, C. W. & Ostendorf, M. (1991). Automatic recognition of prosodic phrases. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Toronto, Canada 1, 321–324.

Wightman, C. W. & Ostendorf, M. (1992). Automatic recognition of intonation features. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, San Fransisco 1, 221–224.