29

Comparing Approaches to Pitch Contour Stylization for Speech Synthesis

Piet Mertens
Frédéric Beaugendre
Christophe R. d'Alessandro

ABSTRACT This chapter describes two approaches to pitch contour stylization as well as a perception experiment to evaluate and compare both methods. The first approach uses an automatic stylization procedure, based on perceptual criteria. It outputs the sequence of audible pitch events (static tones, dynamic tones, complex dynamic tones) in the utterance. Both the tonal perception model and the algorithm are described in some detail. The second approach, known as close-copy stylization, is a manual procedure in which a straight-line approximation of the pitch contour is obtained interactively, by resynthesis of the stylized contour and auditory comparison with the original. A perception experiment using synthetic stimuli with stylized contours was run in order to compare and evaluate both approaches. The stylized contours can hardly be distinguished from the natural contours. Tonal perception stylization gives slightly better results than straight-line stylization.

29.1 Introduction

Several approaches have been proposed for the generation of intonation in speech synthesis systems. The prosodic models used in these systems are quite different; they are defined in terms of: (1) pitch target values derived directly from a phonological representation [Pie81, Hir92]; (2) voice source commands [Ohm67, FK88]; (3) standardized pitch movements, obtained using some stylization of F_0 contours [tHa90]. This chapter focuses on the latter approach. Both in synthesis and analysis (or recognition), the stylized pitch contour is a meaningful level of representation. For synthesis, this is obvious: the stylized pitch controls the synthesizer. But also in the case of analysis, the stylization based on perceptual criteria is a meaningful representation because, as will be shown, it provides a transcription of the prosodic auditory events in the utterance.

The purpose of this chapter is to compare two approaches to pitch contour stylization. To this effect a perception experiment was run in which the subjects had to decide whether the pitch contours of a pair of utterances were identical. These

stimulus pairs contained resynthesized utterances using either the original pitch contour, or a stylized contour obtained by one of the procedures under analysis. The proportion of answers for which the subjects hear no difference between the original and the modified contours provides an estimation of the quality of the stylizations. The subjects were native speakers of French; they were judging French stimuli.

The structure of the chapter is as follows. The remainder of this introduction is concerned with stylization in general: what is stylization, what are the components of a stylization procedure, and how can stylizations be classified and compared? The next two sections each describe one particular stylization strategy. The first is an automatic procedure that simulates tonal perception; it was developed recently [tHa95], and will be described in detail. We also give a concise overview of tonal perception. The second approach is the well-known close-copy stylization, which approximates the pitch contour by a sequence of straight lines [tHa90]. The terms "straight-line stylization" and "pitch movements approach" are used also to refer to the second type of stylization. For each of these approaches, we also mention some already available experimental data. Section 29.4 describes a new experiment aimed at a comparison of the two stylization approaches. Finally, the last section provides a general conclusion.

Stylization is often viewed as a way to reduce the amount of information contained in the fundamental frequency tracing in such a way as to retain only those parts of the pitch curve that have a linguistic function in speech communication, and hence are necessary for the synthesis of prosody. Because still too little is known about which parts of F_0 contours are relevant, and how to determine them, there are several approaches to intonation stylization.

When comparing stylization systems, it is useful to decompose the overall process into three successive components. The first is F_0 determination, as F_0 is a major input to the stylization algorithm. The second component is the actual stylization. The result is a simplified pitch curve, whatever procedure is used to obtain this curve. This component can be followed by a classification step, in which parts of the pitch curve are recognized as instances of discrete units within a particular intonation model. In some systems the last two components are merged within a single step, in particular when the intonation model is seen as the set of (normalized) pitch movements.

Stylization procedures can be classified on the basis of the underlying model. The stylization procedure can be purely mathematical, without any reference to the way the speech signal is processed by the human listener. Most current approaches are of this type. However, when for a given utterance one compares the physical F_0 curve with what one hears, it is obvious that many variations go unnoticed, whether these variations correspond to parts of sounds or to parts of syllables. So what can be heard is only a subset of what is measured. As a result, the stylization can be based on the way in which humans perceive pitch changes in speech signals; that is, it can be based on tonal perception, and in its strongest form stylization is a computer simulation of tonal perception. Although the mathematical and perceptual approaches can be equally successful for speech

synthesis, the latter allows one to gain insight into the process of human perception of prosody.

Another dimension for classification is the criterion used in assessment. When evaluating stylization through comparison of the stylized utterances with their original counterparts, one can use a strong or a weaker criterion; either one verifies whether both versions are indistinguishable, or whether they are functionally equivalent. But these are two completely different questions: in the former case the optimal stylization is the auditory image, in the latter it can be something else, such as the units in the intonation model.

29.2 Automatic Stylization Based on Tonal Perception

This section deals with the stylization procedure based on perceptual criteria. It will be referred to as the automatic tonal perception stylization (ATS). Here we explain the rationale of the approach, quickly introduce concepts of tonal perception, describe the algorithm, and give some of the results obtained with this method.

29.2.1 Rationale

To start, there is an obvious question that needs to be asked: *Why would one make the effort to simulate tonal perception?* The first motivation is due to its theoretical scientific interest. A simulation is a kind of verification procedure. Via listening tasks with stylized stimuli, we can test the accuracy of the perceptual model and modify it where necessary. As the stylization is controlled by two parameters, which are perceptual thresholds (the glissando threshold and the differential glissando threshold; see next subsection), it can be used to measure these thresholds. The stylization thus becomes a tool for basic research on tonal perception. The second motivation is that the stylized contour is an *estimation* of the pitch pattern as perceived by the average human listener, rather than the sequence of the (recognized) language-specific and theory-specific intonation units.[1] The stylized contour thus reflects a representation 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 particular intonation model, or even without reference to the communicative function of pitch in speech! Third, the stylization is independent of any particular linguistic intonation model and can in fact be used to construct such a model in an unbiased way.

[1] It should be noted, however, that the segmentation into syllabic nuclei is to some extent determined by language-specific factors, such as the existence of syllabic consonants in the language. In the current implementation for French, syllables are formed around vocalic nuclei, and the part of the F_0 contour that is analyzed corresponds to the voiced part of the syllable.

29.2.2 Tonal Perception and Prosodic Analysis

Both the peripheral auditory system and the perceptual system shape the speech signal into a mental auditory signal that is quite different from the acoustic signal, to say the least. To illustrate sensory and perceptual processing, one could mention the frequency analysis in the cochlea, the role of frequency response nonlinearity and critical bands for human pitch determination, and so forth. Given these phenomena, it will be clear that (1) the pitch perceived by the human listener does not closely match the fundamental frequency measured in the acoustic signal and (2) the communicative function of prosody can be conveyed only by those pitch events that are preserved in the auditory signal, rather than by any measurable F_0 variation. Consequently, it is useful to obtain the auditory representation. We do not claim that a perceptual model should mimic the peripheral auditory system (although this would, of course, provide the most accurate simulation), but rather that it should take into account the major perceptual effects observed in psychoacoustics.

We briefly describe three perceptual effects related to frequency variations. The first is known as the *glissando threshold*. A fundamental frequency variation that takes place during a given time interval will be perceived as a pitch movement if the rate of change exceeds some minimal amount; this amount depends on the duration of the transition: the shorter the stimulus, the larger the required frequency change. Frequency variations below this threshold are perceived without pitch change (i.e., they are perceived as static tones). The glissando threshold has been measured for pure tones and synthetic vowels generated with a linear frequency change. It should be pointed out that the stimuli used for the determination of the glissando had a constant amplitude. If a glissando threshold for continuous speech could be accurately determined, we would be able to determine which frequency changes are heard as dynamic pitch changes and which as static pitch events.

Of course, the frequency variations observed in actual speech usually exhibit more complex patterns. One observes changes in slope, for instance when a rise is followed by a fall, or when a slow rise changes into a steep rise. Now, let us assume that small slope changes go unnoticed: in this case we can consider the entire frequency variation as a single movement, measure its frequency change, and confront it with the glissando threshold. However, if a given change in slope is audible as such, the variation should be divided into two parts at the point of change in direction, and the frequency change in each part should be evaluated with respect to the glissando threshold. For this reason it is important to know under which conditions a change in slope is perceived. The critical slope change is called the *differential glissando threshold*. There has been very little research on this effect. Note that the proposed procedure still assumes that amplitude changes (as observed in speech) have no effect on tonal perception.

As shown by House [Hou90], the perception of pitch variations is influenced by changes in amplitude and in spectral composition. For instance, a signal with constant fundamental frequency that shows rapid and substantial amplitude dips of some minimal duration will be perceived as a sequence of tones, starting at the amplitude dips. The same holds for signals containing unvoiced parts. Speech

signals are indeed characterized by rapid amplitude changes (e.g., for plosives) and by unvoiced intervals (e.g., unvoiced fricatives). But speech signals are also characterized by slow amplitude changes (e.g., nasal consonants) and progressive transitions from quasi-periodic to aperiodic sounds. A complete quantitative model of this *segmentation effect* is required to deal with those common cases in an appropriate way. Its effect would be to transform the pitch contour at the auditory level into a sequence of short duration tones corresponding to the syllabic nuclei. However, the lack of a quantitative model for this effect makes it difficult to say which parts of the signal make up the syllabic nuclei. As a first approximation, our tonal perception model uses the voiced parts of syllables as the intervals corresponding to the tones.

Automatic perceptual stylization simulates these three perceptual effects. The segmentation effect results in a segmentation of the speech signal into a sequence of short pitch variations. The differential glissando is used to decompose such a pitch variation into uniform pitch movements (rise, fall, level); they are called *tonal segments*. Finally, the glissando threshold determines which tonal segments correspond to audible pitch changes and which are static.

29.2.3 Description of the Algorithm

The stylization procedure consists of several processing steps, some of which are purely acoustic (pitch determination, voicing determination) whereas others are related to perception. We first give an overview of the main processing steps and later describe them in more detail. Figure 29.1 gives a schematic description of the algorithm.

The perceptual model evaluates fundamental frequency variations for syllable-sized fragments of the speech signal. This requires the determination of fundamental frequency and a segmentation of the signal, which, in the current implementation, provides the sequence of voiced portions, one for each syllable in the speech signal.

In the next stage, a short-term perceptual integration (see below, weighted time-average model) is applied to the F0 of each voiced fragment, resulting in a somewhat smoothed pitch contour.

For each voiced portion, the obtained pitch curve is divided into uniform parts (*tonal segments*) on the basis of two perceptual parameters: the glissando threshold and the differential glissando threshold. Ideally, a tonal segment will correspond to a single audible pitch event (rising, falling, or level). Each syllable contains one or more tonal segments, each of which is either static or dynamic (rise or fall). The tonal segment is characterized by the time and pitch of its starting and ending points. The actual stylization is trivial: it consists of a linear interpolation between the start and end points. It will be viewed as an estimation of the perceived pitch (and of the audible pitch movements).

This representation can be further processed within the context of a language-specific intonation grammar in order to go from the level of auditory events to that

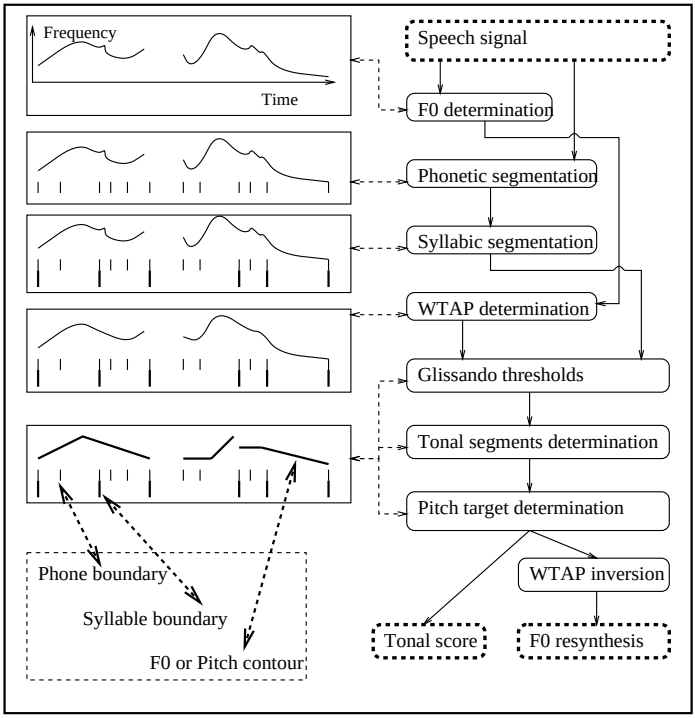

FIGURE 29.1. Automatic intonation analysis algorithm. The left side of the illustration gives a schematic representation of the shape of the pitch contour, in relation to the processing steps, shown on the right side. WTAP stands for weighted time-average pitch.

of the linguistic units [Mer87a]. However, the latter aspect will not be dealt with here.

In what follows, each processing step is described in more detail (see also [dM95]).

1. *Fundamental frequency measurement.* In principle, any kind of pitch determination algorithm can be used, provided its precision is as good as that of human listeners. Most current pitch extractors meet this requirement. Their average accuracy is sufficient. However, results can vary substantially from one algorithm to the other, especially for transitions (unvoiced to voiced, plosive to vowel, glottal stops) and vocal fry. In our implementation the spectral comb method was used as a basic extractor, combined with a postprocessor, which traps many octave shifts.

2. *Voicing determination.* The current implementation uses a simple voiced/unvoiced detection based on energy and zero-crossing rate. Of course, more sophisticated approaches could be used.

3. *Syllabic segmentation.* As said above, the perceptual segmentation effect decomposes the speech signal into a sequence of short tones corresponding to the syllabic nuclei. In the absence of a quantitative model for this effect, the syllabic nucleus is obtained as the voiced portion of the syllable. To avoid artifacts in the resynthesized pitch due to segmentation errors rather than to the stylization itself, an accurate segmentation is required, and for this reason the phonetic labeling provided by the LIMSI speech recognizer [GLAA93] was used. An additional algorithm groups the phonetic segments into syllables. This segmentation is then aligned with the voicing decision in such a way that the sequence of voiced parts is obtained, one per syllable. Another type of segmentation into syllabic nuclei is proposed in [Mer87b].

4. *Short-term pitch integration.* The auditory system seems unable to follow rapid short-term changes in fundamental frequency. There is evidence that an integration process takes place in pitch perception. This phenomenon was observed in a study on vibrato perception [dC94], which proposes a weighted time-average (WTA) model for the perception of short tones. When this WTA model is applied to the F_0 data of each voiced part, a smoother pitch curve is obtained.

5. *Stylization.* Stylization will depend on the settings of two parameters, each of which corresponds to a perceptual threshold: the glissando threshold and the differential glissando threshold.

   The following procedure is applied to each syllable in the utterance, more specifically to the pitch values in the voiced region of those syllables. The syllabic pitch contour is divided into parts with uniform slope, called "tonal segments," in such a way that pitch changes below threshold are normalized to static tonal segments, and that slope changes between two successive tonal segments must be audible (otherwise they should be merged in a single tonal segment). While weighted time average pitch values are used for contour segmentation, F_0 is used for evaluating the frequency changes in relation to the thresholds. The algorithm imposes no limitation whatsoever on the number of tonal segments within one and the same syllable; consequently, any number of pitch movements per syllable are accepted: there can be none (static), one (rise or fall), two (rise-fall, etc.) or more (e.g., rise-fall-rise). The stylized contour is given by the linear interpolation between the WTA pitch values at the boundaries of the tonal segment(s) in the syllable.[2] For static tonal segments the pitch value of the end point is extrapolated throughout the entire segment.

[2]Interpolation is done on a linear frequency scale (Hz), whereas a logarithmic (semitone) scale could have been used (as is the case for close-copy stylizations). There is no evidence in the literature about the perceptual relevance of the differences between the two types of interpolation. A comparison between straight-line stylization and other types of interpolations based on pitch targets [tHa91] shows that these methods were perceptually equivalent.

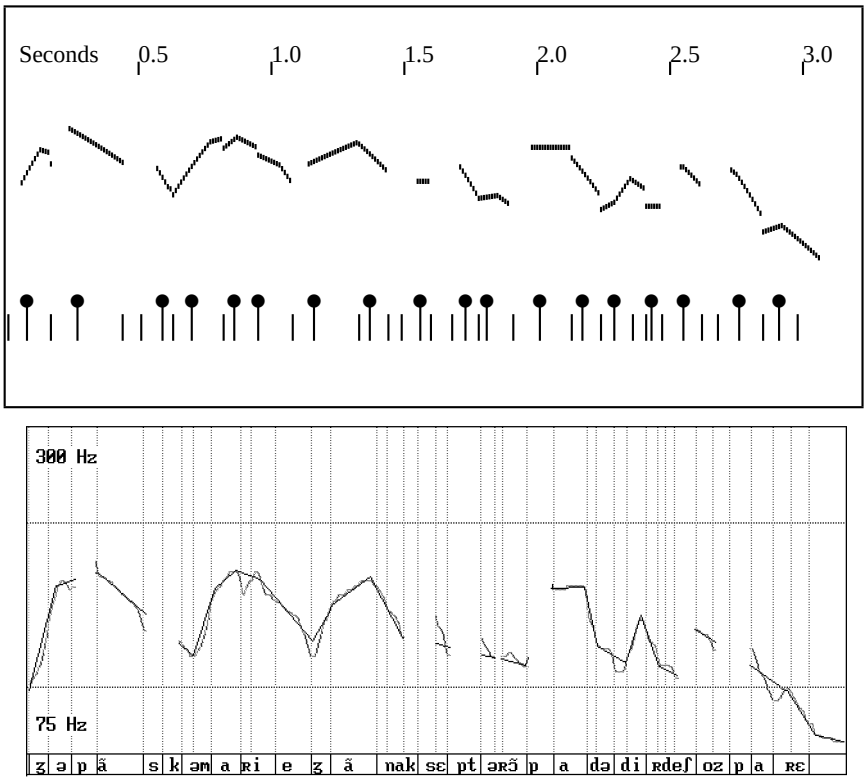

FIGURE 29.2. Automatic tonal perception stylization (top) and straight-line stylization (bottom) for the utterance "Je pense que Marie et Jean n'accepteront pas de dire des choses pareilles." The vertical markers in the upper part indicate boundaries between phonetic segments, and the bullets indicate vowel onsets.

The result of this stylization is referred to as the *tonal score*.

As can be seen in figure 29.2, the tonal score sometimes contains blanks between successive syllables, even though these blanks correspond to voiced parts in the speech signal. This is due to the simple voicing detection algorithm, which also takes into account the energy level.

6. *Resynthesis.* It is possible to reconstruct a synthetic F_0 contour, starting from the stylized pitch contour of the tonal score. In principle this reconstruction is needed because the tonal score is a perceptual representation based on the integrated pitch data, whereas the synthetic speech is an acoustic signal. If the stylized pitch contour were used directly as the pitch for synthetic signal, the perceptual integration would be applied twice: first during the stylization and second by the auditory system of the subject listening to the synthetic signal. The reconstruction of the synthetic F_0 is obtained by passing the tonal score through the inverse of the weighted time-average model.

Figure 29.3 illustrates the pitch contour after different processing steps of the algorithm. The first curve represents F_0, i.e., the output of the pitch determination algorithm. The second curve represents pitch at the output of the weighted time-average model. One can notice that small variations are smoothed. The third curve is the stylized pitch, i.e., the tonal description of the intonation contour. Finally, the last curve is the stylized contour passed through the inverse WTA model; this is the pitch control used in resynthesis.

29.2.4 Discussion

When the model parameters are set to the thresholds as observed in psychoacoustics, the resulting stylization very closely matches the measured pitch (i.e., WTA-pitch) contour.[3] These "standard" thresholds were measured for acoustically simple stimuli (pure tones, synthetic vowels), which are presented in isolation and repeated several times. The thresholds for continuous speech will undoubtedly be higher, because of the acoustic (spectral) complexity and the absence of stimulus repetitions. An important asset of the stylization based on tonal perception is that the stylization itself can be used to measure glissando and differential glissando thresholds for continuous speech. Examples of stylized contours obtained with ATS using different values for the thresholds are presented in *sound example 4* (see [AM95]).

There are, however, some problems that need to be solved first. A major problem is that the approximation of the syllabic nucleus, as the voiced part of the syllable, is inaccurate. New psychoacoustic research is needed to provide a solution. Another problem is the errors introduced by large microprosodic excursions due to unvoiced-voiced coarticulation, in combination with the smearing effect of the WTA model. In order to avoid these errors, a simple microprosody preprocessor (such as described in [Mer87a, Mer89]) can be used. However, it would be preferable to study the perceptual processing of typical microprosodic patterns and to adapt the model for short-term perceptual integration of pitch accordingly.

29.3 Manual Straight-Line Stylization

Manual straight-line stylization (MSLS) is a procedure by which the observed pitch contour is replaced by a less complex contour, having the form of a concatenation of straight lines. It is based on the hypothesis that unnecessary details of the natural melodic curves can be ruled out without any perceptual change. No structural assumption has been made up to now about the nature of such details. For instance, some of the pitch variations related to micromelody (defined here as segmental

[3]The analysis procedure described above was tested in a same-different task using synthetic speech stimuli, based on the original or the stylized pitch contour. This experiment is described in [AM95], and is not repeated here.

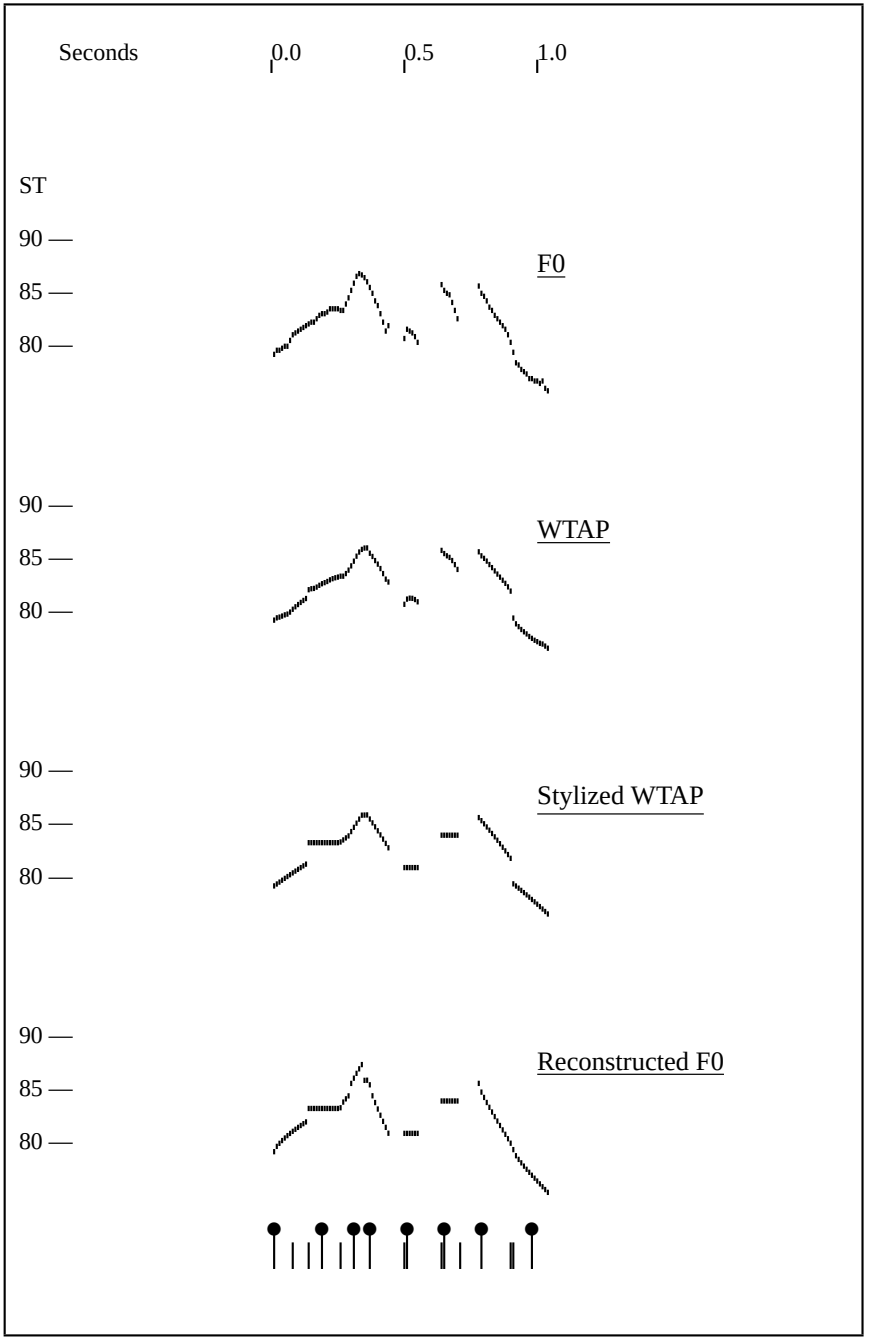

FIGURE 29.3. Pitch curves between processing steps in the stylization algorithm for the utterance "Anne-Marie était effondrée." WTAP is the weighted time-average pitch. Stylized WTAP is the stylized pitch contour. The lower tracing is the pitch contour used for resynthesis.

influences on melody) may be deleted, if they are not perceived, but other details must be preserved. The only aim of stylization is to obtain reduced contours, which must be perceptually identical to the original ones.

An approach to performing such a task was proposed in [tHa90]. In the process of stylization, natural melodic contours are reduced to a concatenation of straight lines on a logarithmic frequency scale (semi-tones/second).

The stylization is obtained interactively by means of an analysis-through-resynthesis technique. The process of stylization is a loop containing three steps: (1) a piece of the pitch contour is replaced by a straight line, according to the (manual) selections by the phonetician; (2) a speech signal with the modified pitch contour is synthesized; and (3) the phonetician listens to the synthetic signal and compares it to the original one. This procedure is repeated for that part of the contour until the original utterance and the modified utterance (i.e., with the stylized contour) are judged equivalent. The same procedure is applied to all parts of the contour.

As a general principle, a minimum number of straight lines is searched for a given contour. These straight lines are called *pitch movements*. The concatenation of pitch movements is called the *stylized pitch contour*. The resynthesized sentences using stylized pitch contours are called *close-copy stylizations*. Again, stylization is only a way to perform data analysis, and no special meaning is associated with pitch movements, so far. It is clear that the data reduction performed provides a better basis for further analyses of pitch contours.[4] Linear predictive coding (LPC) is often used as an analysis/resynthesis technique to obtain close-copy stylizations; but, of course, other techniques can be used as well. *sound example 1* (see [dM95]) presents LPC close-copy stylizations.

This methodology was applied to French in [BdLT92], the aim was to develop a melodic model of French intonation for use in a text-to-speech synthesizer.

While stylization enables a simplified description of intonation contours, without a loss of prosodic information, the degree of abstraction achieved is insufficient to serve as a description of the melodic properties of the language. On the one hand, it is clear that the human auditory system imposes some limits below which two pitch contours cannot be distinguished, and these limits can be determined for each acoustic dimension of prosody (e.g., pitch slope, frequency range, duration of the variation, direction). On the other hand, even if two pitch contours can be differentiated from a perceptual point of view, the prosodic information conveyed by them may still be identical. Taking these constraints into account requires a next stage of data reduction, based on the perceptual equivalence of pitch contours. The data reduction is obtained by a classification of the stylized pitch movements into normalized elementary units. This stage is essential for the development of

[4] A comparison between straight-line stylization and other types of interpolations based on melodic target values is reported in [tHa91]. It appeared that these methods were perceptually equivalent. In particular, the angular points created at transitions between two straight lines do not have any special effect on the resulting melody, as compared to smooth transitions.

a melodic model for TTS synthesis. *Sound example 2* (see [dM95]) illustrates the differences between stylized contours and contours consisting of standardized movements.

As this classification was to be used as the phonetic specification of prosody in a text-to-speech system for French, a set of rules was developed to automatically generate intonation contours from written text, tagged with syntactic information [BdLT92]. Examples of the output of the LIMSI text-To-speech synthesizer using these rules are recorded in *sound example 3* (see [dM95]).

29.4 Comparing Perceptual and Straight-Line Stylizations

This section describes a perception experiment to compare the two types of stylization described above. First some general observations will be made about the underlying assumptions and the results of the two approaches.

29.4.1 Differences Between the Two Approaches

Straight-line stylization is a manual procedure, whereas tonal perception stylization is automatic. For the sake of comparison, we will assume that close-copy stylization can be obtained automatically; indeed such an algorithm has already been proposed [SDHG93].

Both stylization algorithms have some characteristics in common. They both take into account amplitude variation and voicing, although in a different way. In the case of ATS this is done explicitly in the phonetic segmentation step. In the case of automatic close-copy stylization it is implicit in the weighting of F0 data points. Both approaches need a way of handling microprosodic perturbations. This aspect can be integrated in ATS because the phonetic segmentation provides the identification of the phonetic segments. Automatic straight-line stylization uses vowel onset detection.

A major difference between ATS and MSLS is the scope of pitch movements. In ATS, the syllable was chosen as the basic intonation unit (syllabic tones) at the linguistic level, although at the auditory level a syllable can contain multiple tonal segments, of course. This was motivated by the perceptual relevance of syllables for the segmentation of pitch contours. By contrast, straight-line stylization is based on units (pitch movements) that may encompass several syllables, or only a part of a syllable. This effect is visible in figure 29.2. Generally, MSLS is more global: it represents larger intonation units (containing several syllables). The same pitch movement can group a series of static or dynamic tones. The three tones at the end of *n'accepterons* have clearly no individual linguistic significance, and it is simpler in this case to group them in a single pitch movement.

It should be pointed out, however, that the linguistic intonation model [Mer87a] underlying the ATS approach also defines units ranging over several (unstressed)

syllables as well as prosodic groups comprising both stressed syllables and sequences of (one or more) unstressed syllables. These larger units are supposed to be formed at a higher level of perceptual processing and are language-specific. A description of the great variety of pitch contours observed in spontaneous speech would require a very large amount of basic pitch patterns (corresponding to prosodic groups); it would require a smaller number of units if pitch movements (defined in terms of the pitch variation) were used; however, it can be described most economically as the combination of syllable-sized components.

The pitch movement approach seems particularly efficient in terms of simplicity of intonation rule design. But a drawback of the approach is that the pitch movement inventory was designed to model a given speech corpus, and it is not clear whether the set of pitch movements obtained can actually be used to synthesize pitch contours in any speaking style.

For these reasons, we think that the grouping of syllabic tones within the same pitch movement should be done at a higher level. It is not a matter of stylization, because it is dependent on several factors such as stress and the intonation grammar of the particular language.

29.4.2 Perception Experiment

An experiment was conducted in which the two types of stylization were presented to the same group of subjects for comparison.

Stimuli

For this experiment, 20 sentences were selected from a speech database of 60 sentences read by one male speaker of Parisian French, taking into account some syntactic, phonotactic and lexical constraints. All sentences were relatively short (between two and eight words) in order to avoid problems of short-term auditory memory when comparing the two versions (original and stylized) of the sentence. All stimuli were generated using LPC resynthesis. For each sentence, the stimulus groups are labeled V1, V2A, V2B, V3, and V4. V1 is the resynthesized original signal, V2A corresponds to the ATS stylization, V2B is the close-copy stylization (MSLS), and V3 and V4 are two alternative versions of the MSLS, with pitch contours that are increasingly different from the original contour. They were included in the test material in order to obtain a range of pitch contours going from identical, over almost identical, to clearly different.

For category V1, LPC-resynthesized sentences (with the original pitch contour) were used rather than the original sentences because of the quality degradation introduced by LPC: the quality difference between an original sentence and the corresponding LPC stylized version would have been easier to detect than the difference in intonation.

In the V3 and V4 categories, the alternative versions were derived from manually stylized contours (i.e., from V2B) in which either the slope, the timing, or the frequency level of a movement (of the overall close-copy contour) had been modified. A change in slope will affect the duration of the pitch movement; the

timing of the start of the movement was modified such that the end time of the movement remained unchanged. For these modifications, we referred to perceptual experiments on the differential thresholds of pitch and pitch change [IS70, tHa81]. The modifications of slopes and levels chosen for categories V3 and V4 were of the order of 1.5 and 3 times the thresholds of "just noticeable difference," respectively. (For more details, see [Bea94], p. 65, table 2.1.)

Procedure

Subjects were asked to concentrate on intonation alone, and not on other aspects of the signal. For each pair, they had to indicate whether the two stimuli in the pair were identical with respect to intonation.

The subject sat in front of a computer with a mouse device. Each stimulus pair was presented once, after which the subject had to enter his response ("same" or "different") by clicking in the appropriate box on the screen.

Stimuli of the five conditions (V1-V1, V1-V2A, V1-V2B, V1-V3, V1-V4) were presented in random order. The order of the two stimuli within a pair was also varied randomly between X-Y and Y-X (e.g., V1-V3 and V3-V1).

Subjects

There were 10 subjects. None of them had known hearing loss. Tonal audiograms were made before the experiment to verify this. One of the authors also participated in the experiment as a subject (d'Alessandro).

Results

Table 29.1 summarizes the results of this experiment. About 93% of pairs in category V1-V1 were perceived as identical (100% were identical), and about 90% of pairs in category V1-V2A and 88% of pairs in category V1-V2B were also perceived as identical. The scores were only about 50% and 24% for categories V1-V3 and V1-V4, where the difference between the two parts of the stimulus pair becomes progressively larger. The scores thus follow the expected trend: the larger the difference between the two parts of the stimulus pair, the lower the proportion of "same" answers. This indicates that the subjects were able to perform the task, to perceive modifications in the pitch contours.

The fact that we did not obtain 100% "same" ratings for the V1-V1 category is normal in perceptual tests and can be ascribed to unsystematic errors of judgment and variation in the subjects' levels of attention.

The mean difference between the ratings for identical stimuli (V1-V1) and those for pairs containing the automatic stylization (V1-V2A) is only about 3%. The mean difference between the V1-V1 category and the manual stylization (V1-V2B) is only about 5%. This might indicate that even if a slight difference between original and stylized contours exists, it can be ignored, and the stylized contours can thus be considered perceptually equal to the original ones.

We can conclude that the two types of stylization (MSLS and ATS) give fairly similar results in terms of the perceptual equivalence between stylized and natural contours.

On the average, tonal perception stylization gives slightly better results than straight-line stylization. However, the results are very close: the mean difference

TABLE 29.1. Perceptual ratings for manual straight-line stylization and automatic tonal perception stylization. The columns are the subject identification, the total number of stimulus pairs in the test set (NSP), followed by the proportion of stimulus pairs judged identical for the five stimulus types. V1 is the resynthesized original utterance, V2A is the synthetic utterance with the pitch contour obtained with the automatic tonal perception stylization, V2B is the synthetic utterance with the pitch contour obtained using the manual straight-line stylization, and V3 and V4 are alternative versions of the manual stylization.

Subject	NSP	V1V1	V1V2A	V1V2B	V1V3	V1V4
JKas	348	82.8	82.3	67.6	38.1	19.1
XLap	449	89.4	90.4	86.9	71.4	52.8
CdAl	567	92.9	82.3	86.7	40.1	3.1
BDov	449	94.3	92.7	84.5	29.9	8.4
SRos	453	94.1	95.4	85.2	47.7	22.2
ABra	456	100	90.9	97.7	53.9	13.2
TLeb	467	95.3	97.0	97.9	53.0	27.8
LLac	452	89.7	86.5	90.3	58.6	24.8
SBen	575	98.8	88.7	95.2	42.3	28.9
MJar	455	94.1	93.6	85.0	63.3	41.0
ALL	4681	93.1	90.0	87.7	49.8	24.1

between the V1-V2A and V1-V2B categories is only about 2%. It should be pointed out that 4 subjects out of 10 rated straight-line stylization higher than tonal perception stylization. The fact that ATS gives better overall results than MSLS is somewhat surprising because the latter uses an interactive procedure in which audible differences will be eliminated as much as possible during the stylization procedure, thanks to the auditory feedback. A possible explanation for the good results for ATS is that it contains more line segments than straight-line stylization, resulting in a better match with the original pitch contour. Examples of stylized contours obtained with ATS using different values for the thresholds are presented in *sound example 4*.[5]

29.5 Conclusion

The experiment described in this chapter demonstrates that the automatic stylization of pitch contours based on tonal perception produces a simplified contour that is hardly distinguishable from the original, and is as good as, or even better than, the stylization obtained with the manual, interactive procedure known as close-copy stylization.

[5]These examples correspond to stimuli V1, V2, V3, V4 in [dM95].

Pitch contour stylization in general is a powerful tool for designing prosodic models in speech synthesis. Such a model was built for French, according to the close-copy stylization (or pitch movement, or straight-line) methodology [Bea94]. It is integrated within the LIMSI text-to-speech system. Two types of problems with this approach to stylization were encountered. On the one hand, the stylization process is time consuming, and the experimenter has to make ad hoc decisions regarding the relevant features and movements that are needed. On the other hand, the pitch movements obtained could be dependent on the specific characteristics of the speech corpus used and are not strongly linguistically motivated.

Therefore, it was decided to design another type of intonation stylization to overcome the above-mentioned problems. This stylization procedure is automatic and grounded on perception. Because it is automatic, the procedure is fast and efficient. Because it is grounded on perception, the procedure separates the linguistic and perceptual-acoustic aspects of F_0 contours.

As for the perceptual equivalence between stylized and natural F_0 contours, the two types of stylization processes seem almost comparable. This means that F_0 contour stylization is not unique, but is dependent on the underlying perceptual and linguistic assumptions.

The pitch movements approach is an efficient representation for designing intonation synthesis rules, with the above-mentioned limitations in mind. Automatic tonal stylization represents intonation at a lower level of description. Therefore, it should make a better framework for intonation rule writing, but more rules would be needed. ATS makes no assumptions on what is relevant and what is not in the processing of prosodic features by the human listener; it merely applies the findings of psychoacoustics. It could also be used for further automatic processing of intonation (such as automatic transcription of prosody) and for computer assisted teaching of intonation.

Acknowledgments: The authors would like to thank the subjects for their kind help in the course of this research, as well as the two anonymous reviewers for their valuable comments.

References

[BdLT92] F. Beaugendre, C. d'Alessandro, A. Lacheret-Dujour, and J. Terken. A perceptual study of French intonation. In *Proceedings of ICSLP'92*, Banff, Alberta, Canada, 739–742, 1992.

[Bea94] F. Beaugendre. *Une étude Perceptive de l'Intonation du Français.* Doctoral dissertation, LIMSI report NDL 94-25, Université de Paris-sud, Orsay, 1994.

[dC94] C. d'Alessandro and M. Castellengo. The pitch of short-duration vibrato tones. *J. Acoust. Soc. Amer.* 95(3):1617–1630, 1994.

[dM95] C. d'Alessandro and P. Mertens. Automatic pitch contour stylization using a model of tonal perception. *Comp. Speech and Lang.* 9:257–288, 1995.

[FK88] H. Fujisaki and H. Kawai. Realization of linguistic information in the voice fundamental frequency contour of the spoken Japanese. In *Proceedings of IEEE-ICASSP*, New York, 663–666, 1988.

[GLAA93] J. L. Gauvain, L. F. Lamel, G. Adda, and M. Adda-Decker. Speaker independent continuous speech dictation. In *Proceedings of Eurospeech'93*, Berlin, 125–128, 1993.

[Hir92] D. J. Hirst. Prediction of prosody: An overview. In *Talking Machines: Theories, Models, and Designs*, G. Bailly and C. Benoit, eds. North-Holland, Amsterdam, 1992.

[Hou90] D. House. *Tonal Perception in Speech*, Lund University Press, Lund, Sweden, 1990.

[IS70] A.V. Issachenko and H.-J. Schädlich. *A Model of Standard German Intonation.* Mouton, The Hague, Paris, 1970.

[Mer87a] P. Mertens. L'intonation du Francais. De la description linguistique à la reconnaissance automatique. Unpublished doctoral dissertation, Catholic University of Leuven, 1987.

[Mer87b] P. Mertens. Automatic segmentation of speech into syllables. In *Proceedings of the European Conference on Speech Technology*, Edinburgh, UK, 9–12, 1987.

[Mer89] P. Mertens. Automatic recognition of intonation in French and Dutch. In *Proceedings of Eurospeech 89*, Paris, France, 1:46–50, 1989.

[Ohm67] S. Ohman. Word and sentence intonation: A quantitative model. *K.T.H. Quarterly Progress and Status Report*, 2:20–54, 1967.

[Pie81] J. Pierrehumbert. Synthesizing intonation. *J. Acoust. Soc. Amer.* 70:985–995, 1981.

[Spa93] G. W. G. Spaai, A. Storm, A. S. Derksen, D. J. Hermes, and E. F. Gigi. *An Intonation Meter for Teaching Intonation to Profoundly Deaf Persons.* IPO Manuscript no. 968, Inst. for Percept. Res., Eindhoven, 1993.

[tHa81] J. 't Hart. Differential sensitivity to pitch distance, particularly in speech. *J. Acoust. Soc. Amer.* 69(3):811–821, 1981.

[tHa90] J. 't Hart, R. Collier, and A. Cohen. *A Perceptual Study of Intonation.* Cambridge University Press, Cambridge, 1990.

[tHa91] J. 't Hart. F_0 stylization in speech: Straight lines versus parabolas. *J. Acoust. Soc. Amer.* 90(6):3368–3370, 1991.

Appendix: Audio Demos

A sound demonstration is provided.