Automatic Intonation Extraction and Generation for French

Eric Keller and Stefan Werner
LAIP-Lettres, University of Lausanne, CH-1015 Lausanne, Switzerland
and Linguistics and Phonetics, Joensuu University, FIN-80101 Joensuu, Finland

Abstract

It is now possible to develop interactive and automatized intonation training modules for personal computers. The initial challenge is to provide a pitch-extraction technique that is fast and reliable. A greater challenge is to create a good intonation model for a given language, to permit automatic generation of model pitch contours. The first author has implemented a number of pitch extractions in his software Signalyze and is cooperating with the second author on a project for a high-quality synthesis of French. Different techniques of pitch extraction and pitch generation techniques will be discussed and demonstrated.

Keywords

Intonation, French, Pitch extraction, Pitch generation, Speech synthesis

Introduction

The acquisition of speech proficiency in a foreign language can be facilitated by visual feedback on specific aspects of the target language. Intonation is one aspect that can be reliably extracted from the speech signal, and that can be presented in an easily understandable form. In conjunction with appropriate timing information, fundamental frequency contours furnish most of the significant prosodic information contained in speech expressions. It is thus at the center of much current interest in speech analysis and speech modeling.

In our work for the speech analysis program Signalyze and our new high-quality speech synthesizer LAIPTTS, we have studied pitch extraction and pitch generation extensively. We have developed a pitch extraction algorithm for Signalyze, and we have implemented a pitch generator in our speech synthesizer for French. The underlying concepts for the two types of development will be discussed here.

Fundamental Frequency Extraction

With respect to pitch extraction, it is useful to contrast the advantages and disadvantages of algorithms based on Fast Fourier Transforms (FFTs), autocorrelation and temporal structure analysis (TSA) (for a review of pitch extraction techniques, see Hermes, 1993, and Hess, 1983; see Keller, 1994, for the Signalyze implementation of discussed algorithms).

FFT-based pitch extraction uses information from both the fundamental frequency (F0) and its harmonics (H1...x), and is thus quite reliable. However, its precision is a function of the number of points over which the FFT is calculated. A fine-grained frequency precision requires a relatively wide temporal window, and the width of this window may in turn interfere with the precision required along the time axis. While FFT-based pitch extractors may be relatively robust, their frequency-time precision trade-off, as well as their time requirements on the common desktop computers of today, thus make them less than ideal choices for didactic feedback devices.

Autocorrelation is another robust algorithm, since it exploits all possible statistical relationships between successive fundamental pitch periods. The algorithm derives its frequency information from the time period separating two successive and mathematically similar pitch periods. Correlations are calculated between the numeric values describing a given pitch period, and sets of values derived from successive time windows of similar length. The point of greatest similarity is identified, and the delay between the two similar time windows is converted into an instantaneous fundamental frequency. When used with the sampling frequencies for speech in common use today (e.g. 12-22 kHz), this routine is quite precise and fairly robust. However, as used on today's common desktop computers, the routine is also too slow for easy didactic usage.

Temporal structure analysis (TSA) concentrates only on identifying the major peaks or valleys that are characteristic of fundamental frequency periods. If the algorithm is well-designed, it can rapidly and quite reliably identify maximum extensions of pitch period contours. Errors occur in three typical locations: when peaks and valleys are ill-defined, when they are overlaid by noise, and when peaks or valleys end in a double-peaked formation. In normal usage under reasonable conditions of environmental noise, these occurrences are sufficiently rare to be only of minor annoyance. Consequently, TSA can be an excellent basis for language-training work on intonation.

Fundamental Frequency Generation

The problem of generating pitch contours is profoundly different from that of extracting F0 contours. Here, the challenge lies in knowing where pitch must rise and fall, and for how long it can remain at a given level. The potential for didactic applications is considerable: If reasonable pitch contours can be created from arbitrary text in a given language, such contours can be used in conjunction with speech synthesis devices for the generation of artificial, but believable speech output on the basis of simple text material.

Quite a few predictive techniques have been developed over the last few years for the purpose of generating pitch contours. They vary considerably in their complexity, in their appropriateness for didactic use, as well as in their suitability for different languages. For simplicity's sake, we shall distinguish just four major types of technique here, (1) phonological prediction schemes, (2) perceptually-derived F0 patterns, (3) neural-network derived F0 patterns, and (4) physiologically-inspired mathematical modelling of F0 contours. Since this is one of the most active areas of linguistic and phonetic modelling today, no attempt is made to provide a complete survey. Many further techniques as well as various hybrid solutions are variously promoted.

For a language like English, a *phonological approach* such as Pierrehumbert (1980, 1981) appears promising. Intonation is represented as s sequence of *low* and *high* tones that can play specific roles such as "pitch accents", "phrase accents" and "junction accents" (found at phrase boundaries). These units are translated into concrete phonetic representations on the basis

of context-sensitive rules that are passed over the underlying phonological structure from left to right. In addition, there are "downstep rules" that combine with a downwards slant of the F0 base line to produce the typical frequency drop-off from the start to the end of a prosodic phrase.

One advantage of this model is that relatively few formal symbols are needed to represent intonation in this system. At the same time, this is also one of the system's limitations, since the speech signal generated on the basis of such a simple system shows many auditorily significant differences from normal speech signals. For one thing, many normal F0 distinctions are not simply high-low distinctions, but depend on the presence on intermediate values. Furthermore, each syllable has a characteristic F0 low-high-low contour, called "microprosody", which is simply left aside in such a predictive scheme. Finally, a phonological predictive scheme that may be adequate for a language like English may not always work in other languages. For example, French is more adequately described as a syllable-based language than as an accent-based language. For this reason, an accent-based predictive scheme for intonation contours is suspect in French (Zellner, 1996).

The next two pitch generation schemes are representative of empirical approaches to the question. Here, the effort focuses on the identification of typical and/or minimal intonational contours that characterize given speech sequences. Once identified and associated with identifiable syntactic and phonological sequences, these patterns can be stored and made available for sequential implementation in a pitch contour.

The intonation model developed at the Dutch Institute for Perception Research (IPO) (see review in 't Hart et al., 1990) aims to incorporate only the *perceptually relevant elements* of an intonational contour. A model for a given language is based on an extensive series of perceptual evaluations of successively more stylized F0 patterns. Once identified, perceptually distinctive patterns are stored in terms of three parameters, which are: their start level, their end level and their steepness. For reinsertion into an F0 contour, patterns are aligned on a sentence by means of guidelines that slope downwards from left to right.

There are two major difficulties with this approach. For one, this is clearly a time-consuming and labour-expensive approach to the problem of identifying the F0 contours of a language. This type of effort is thus not easily undertaken to create the intonational data base for a new language. Secondly, it is possible that the technique suppresses information that may not be judged to be perceptually distinctive, but that in fact helps to improve the quality, and thus the degree of naturalness of a synthesized stretch of speech. Microprosodic modulations, for example, are not captured by this approach; nevertheless, their modelling appears to contribute to improved naturalness.

A less labour-intensive means of obtaining F0 patterns was described by Traber (1991, 1992). In this approach, regular relationships are identified by means of a *neural network* between German sentences, described in terms of syllable-specific features, and their F0 contours. It may be interesting to consider the features that were retained for the prediction of the F0 contour. They were: the syllable's accent, the type of phrase containing the syllable, the presence of a lexical boundary, the presence of a sentence-final condition, and the position of the syllable with respect to the main accent. Further features captured the length of the syllable's vowel, high or low intrinsic F0 values, and the voiced/unvoiced quality of the surrounding consonantal context. Finally, the model was rendered sensitive to immediately preceding F0 conditions.

Both statistical and perceptual results were considered satisfactory, though Traber observes that rare patterns are sometimes ill-identified by the neural network technique. The approach is certainly intriguing. All of the required predictive features can be obtained mechanically from input text. The technique is straightforward and may well be applicable to

new languages without too much difficulty, as long as adjustments are made to compensate for differences in the use of stress.

The final approach considered here is the *physiologically-based model* developed by Fujisaki and his colleagues (see e.g., Fujisaki & Hirose, 1982). This model captures two main contours observed to interact in a large number of languages, a global intonational low-high-low contour extending over the entire prosodic phrase and a set of syllable-wide low-high-low contours. Both types of contour are considered to have a physiological basis in typical expiratory speech patterns. The syllable-wide contours are most prominent during stressed syllables in languages (like English) that show extensive stress modulations; they are in evidence to various degrees during every syllable in those languages (like French) that show less stress modulation. Syllable-level contours are superimposed on the phrase-level contours to produce a complete approximation to the original contour. To calculate an intonational contour in this fashion, only a few speaker-specific and language-specific parameters are required. These can be obtained by effecting a set of successive approximations between the simulated and the natural contours (Werner, 1995).

Despite its simplicity, the model makes reasonably good approximations for the declarative sentences of a number of languages. In our implementation of the Fujisaki algorithm for French, we diverged from the implementation that had previously been used for Japanese in the sense that we generate syllable-level contours for every syllable of a phrase, rather than just for stressed syllables. In this fashion, we generate an approximation to the microprosodic F0 structure, in addition to producing F0 contours that are approximations to those of natural sentences. These contours can be generated in real-time on contemporary desktop computers.

Conclusion

In summary, we estimate that our tools for extracting and generating F0 contours are in the process of attaining an interesting level of maturity. Their use in didactic devices can now profitably be explored. Many commercially available F0 extraction devices are already available. In view of their increasing ease of prediction, F0 contour generation on the basis of text is likely to become available soon for didactic purposes.

References

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

Fujisaki, H., & Hirose, K. (1982). Modelling the dynamic characteristics of voice fundamental frequency with applications to analysis and synthesis of intonation. In *Preprints of the Working Group on Intonation, 13th Intl. Congress of Linguists* (pp. 57-70). Tokio.

Hermes, D. J. (1993). Pitch Analysis. In M. Cooke, S. Beet, & M. Crawford (eds.), *Visual Representations of Speech Signals*, pp. 3-25. New York: John Wiley & Sons.

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

Keller, E. (1994). *Signalyze: Signal Analysis for Speech and Sound (User's Manual).* Lausanne: InfoSignal Inc.

Pierrehumbert, J. (1981). Synthesizing intonation. *Journal of the Acoustical Society of America*, *70*, 985-995.

Pierrehumbert, J. B. (1980). The phonology and phonetics of English intonation. *PhD thesis, MIT [distr. by the Indiana University Linguistics Club]*.

Traber, C. (1991) F0 generation with a data base of natural F0 patterns and with a neural network. *Proceedings of Eurospeech 1991*, 141-144.

Traber, C. (1992) F0 generation with a data base of natural F0 patterns and with a neural network. In: Bailey, Benoit & Sawallis (eds.): *Talking Machines: Theories, Models, and Designs (pp. 287-304).* Elsevier.

Werner, S. (1995). Use of a neural network for parameter optimization in Fujisaki models of intonation. Poster presented at *NODALIDA 95 (Nordic Conference for Computational Linguistics)*, Helsinki, July 29.

Zellner, B. (1996). Structures temporelles et structures prosodiques en français lu. *Revue Française de Linguistique Appliquée: La communication parlée.* 1. (pp.7-23). Paris.