Aneto: a tool for prosody analysis of speech

Miquel Febrer, Albert Febrer, Antonio Bonafonte, Ignasi Esquerra
Universitat Politècnica de Catalunya C/Jordi Girona 1-3 08034 Barcelona, SPAIN
http://gps-tsc.upc.es/veu {febrer|antonio|ignasi}@gps.tsc.upc.es

Abstract

The developed tool provides utilities for prosody analysis and labeling of voice signals. It works under Windows 95 and Windows NT environments and uses the Microsoft Win32 application programming interface (API) for audio playing and recording. The application detects the prosody of speech signal and then the original intonation can be stylized in order to observe the pitch contour. Besides, the original intonation can be easily modified and it is possible to resynthesize the voice signal according to the new intonation. Listening to the resynthesized signal, the user can evaluate the results of the prosodic modification.

1 Introduction

This application aspires to be a helpful tool for prosody analysis and database labeling in the context of the development of the Text-to-Speech (TTS) system that is being developed at the Universitat Politècnica de Catalunya (UPC) [1][2].

The UPC-TTS system is a bilingual system able to read text in Spanish and Catalan that works using concatenative synthesis. The main modules are phonetic transcription, prosody generation and speech generation.

Synthetic speech is achieved by concatenating segments of real speech. All possible combinations of two phones for a given language, called diphones, are stored in the unit database. Each unit has a particular prosody depending on the original recording context from which it was extracted. The function of the speech generation module is to adapt the prosody of these units to the values assigned in the previous module.

However, one of the main problems in speech synthesis is how to determine a model for achieving a natural intonation and rhythm. Analyzing prosody of real speech signals is very useful to test and validate prosodic models for TTS systems.

Aneto is a software application that works under Windows 95 and Windows NT environments. It uses the Microsoft Win32 application programming interface (API) for audio playing and recording. Speech files can be opened, visualized and resynthesized with this application. Fundamental frequency can be calculated and modified using a graphical interface. Labeling of segments in the speech signal is also possible. The tool can be downloaded from [2].

The paper makes an introduction to prosody in next section. In section 3, it details the methods involved in the generation of the pitch contour: pitch detection and error correction algorithms. Pitch contour stylization is detailed in section 4, as well as the speech resynthesis method used to evaluate the suitability of this stylization or the modifications introduced to the original prosody.

2 What is Prosody?

Prosody gives naturalness and message intelligibility to speech. The great importance and complexity of prosody in speech makes this subject an important area of research in speech synthesis applications.

Prosody is the combination of voice's pitch, duration and energy variation during speech. It provides an additional sense to the words, which is extraordinarily important in natural speech. For example, interrogative and declarative sentences have very different prosody (especially intonation). Besides, the prosody of a sentence is one of the factors that make a speaker seem happy, sad, angry or frightened. We can even decide from prosody if the speaker is an energetic person or, on the contrary, a lazy one. When singing, intonation and timing evolution characterize melody.

But prosody is not only important in natural speech but also in synthetic speech. Prosody is crucial in order to achieve an acceptable naturalness. If a TTS system does not have a good prosodic treatment, its output speech sounds completely monotonous and, moreover, it won't be able to distinguish between sentences of different kinds.

In this section, an introduction to intonation and duration as the two main parameters of prosody is given.

2.1 Intonation

In a first approximation, sounds can be classified in voiced and unvoiced sounds. Voiced sounds, unlike unvoiced ones, are produced making the vocal chords vibrate. These vibrations provoke some periodicities in the speech signal and therefore a fundamental frequency (F0). This value is inversely proportional to the distance between periodicities and it makes speech sound with higher or lower frequency. It is commonly called pitch. On the contrary, as unvoiced sounds do not have any periodicities (vocal chords do not vibrate) and can be modeled as a filtered noise signal. So if we detect the pitch curve of a speech signal it will only exist in the voiced segments.

Pitch is not constant but its value changes during a sentence. That is called intonation. Thanks to intonation we can distinguish, for example, between a declarative and an interrogative sentence or identify focused words inside a sentence.

2.2 Duration

The duration of speech segments is the other main parameter of prosody. The timing structure of a sentence is extremely important to give naturalness to speech. Phone duration depends on a great number of parameters, such as its phonetic identity, surrounding phones, level of stress or position in the sentence or in the word. What's more, duration of a word also depends on its importance in the sentence. For example, a focused word will generally have longer duration.

3 Prosody analysis

TTS systems generate speech from a text. There is a need of prosodic assignment to phones to produce high quality speech. Once the phones to synthesize are determined, it is necessary to know the pitch and duration required to generate adequate speech. However, prosodic treatment of TTS systems has not yet achieved the required quality needed for most applications.

This is the main reason why the tool was created: the prosodic module requires models of prosodic patterns and these patterns have to be studied and tested before the application to TTS.

The goal of the prosodic analysis is the generation of the pitch contour. The pitch detection method used in Aneto is detailed in next section.

3.1 Pitch detection

An epoch detection algorithm [3] identifies voiced and unvoiced segments. The algorithm detects glottal closure instants using the Frobenius norm, a morphological filter and a peak detector.

The algorithm obtains pitch marks placed in glottal closure instants. The pitch value (F0) is simply the distance between contiguous marks.

In *Figure 1*, a voiced speech segment is marked with the positions of glottal closure instants.

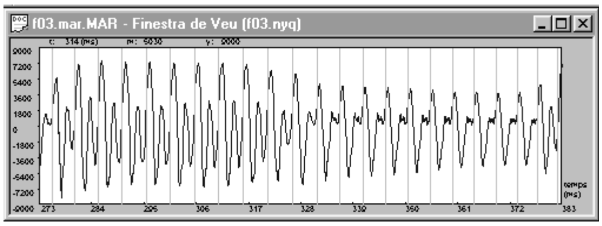

Figure 1. Voiced speech segment with pitch marks obtained by the automatic pitch detection algorithm.

3.2 Error correction

The pitch detection algorithm works reasonably well. However, as many other algorithms, it makes some errors in positioning marks. These mistakes can be corrected manually, dragging a pitch mark with the mouse to the correct position. It is also possible to suppress or add a pitch mark in an arbitrary position.

Some of these mistakes can be typified. Sometimes the algorithm does not put a mark where there should be one, giving a half frequency value in the pitch contour. By the other hand, sometimes it puts an additional mark between two correct marks giving a double frequency value. This is a common problem found in other systems and algorithms. Error correction methods proposed in [4] have been evaluated.

An algorithm that corrects these mistakes has been proposed to improve pitch mark labeling and it can be optionally used in the automatic pitch detection process. It assumes that variations inside a voiced segment should not exceed a 75% of difference between contiguous pitch marks. Therefore variations exceeding this bound are considered erroneous. The algorithm deletes a mark when it detects a pitch value that is more or less the double of the previous one and puts a mark when it is the half, always considering values of surrounding pitch values.

3.3 The pitch contour

Once pitch marks are correctly situated we can find the pitch contour. A pitch value can be assigned to every mark as the distance with the previous one. Pitch values are a good approximation of the voice signal's fundamental frequency. The sequence of pitch values generates a pitch contour, which represents the voice signal's intonation. Obviously the pitch contour does not exist in unvoiced segments.

Optionally, the pitch contour can be smoothened in order to achieve a more uniform presentation and an easier stylization. The post-filtering consists of a median filter and a posterior mean filter.

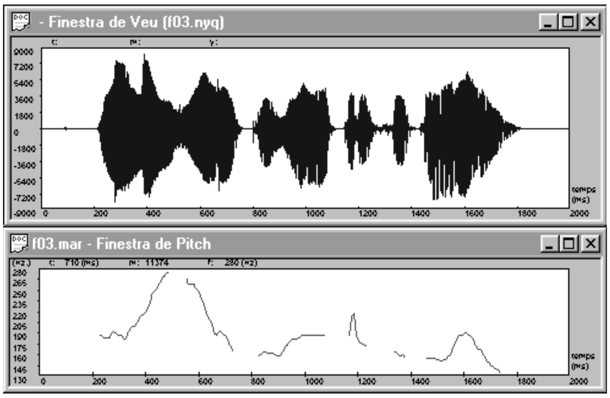

Figure 2. Speech segment and its pitch contour.

An example of a pitch contour can be found in *Figure 2*. Pitch detection has been applied automatically with the error correction algorithms. The curve has been smoothened with post-filtering. The sentence of the example is composed of 6 voiced segments. Unvoiced sounds are the discontinuities in the pitch contour.

4 Pitch contour stylization

Pitch analysis generates complex pitch curves. But intonation is often perceived as the tendencies and inflections of the curve and not by the fine variations. In this sense pitch contour is often stylized to facilitate the observation of its evolution and to simplify the definition of prosodic patterns.

The stylization is automatic and allows manual supervision. Once the stylization is done, the application allows speech resynthesis using the new stylization, either automatic or manual, to test its effects on speech.

4.1 Stylization by linear segments

Prosodic patterns are very often defined by successive linear segments with inflection points tied to parts of a sentence. In order to study pitch evolutions, or to generate prosodic patterns, stylizing the pitch contour can be very useful.

In this application, pitch contour is approximated with a variable number of linear segments. So segments with a flat pitch contour will require a little number of lines in order to achieve an acceptable approximation and irregular curves will require a larger number of lines for similar approximation error. The user can choose the maximum number of lines and the maximum mean error in a voiced segment. The algorithm approximates every segment with the minimum number of lines that achieve the specified error. If the error is not achieved the number of lines is the maximum number of lines. Different straight lines will connect in inflection points.

In *Figure 3* a speech segment is automatically stylized using 3 segments, which is enough to ensure a mean error value smaller than the error bound specified. If the error bound decreased, the third segment would be divided in two lines at least.

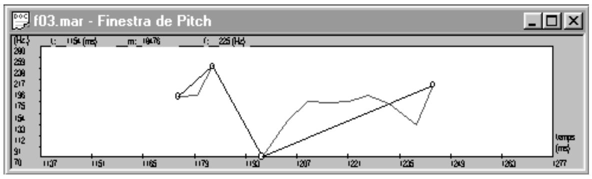

Figure 3. Speech segment stylized with 3 linear segments.

4.2 Intonation and duration modification

Pitch stylization can be modified in order to change the original prosodic pattern to better adjust or test modifications in resynthesized speech. To change a pitch value, the inflection point must be moved vertically, dragging it with the mouse cursor. To modify the original duration of a segment, the suitable inflection point must be dragged horizontally.

In *Figure 4*, a sentence is analyzed. The pitch detection algorithm detects 7 voiced segments. The stylization is performed in different number of lines per segment (for example, 4 lines in the fourth voiced segment and only 1 line in the second voiced segment).

The stylization of the fifth segment is altered to obtain a flat intonation in resynthesized speech. Last voiced segment has been lengthened. The effects can be perceived in the synthetic speech waveform below.

4.3 Resynthesis

The stylization of pitch contour can be evaluated by listening to synthetic speech obtained by resynthesis. The modifications of the pitch contour can also be used to test the effects of pitch and duration alteration on a sentence.

The application includes a synthesizer that can modify pitch and duration of input speech. The synthesis is based on TD-PSOLA (Pitch-Synchronous Overlapp and Add in Time Domain) [5]. This method is used in the TTS system developed at the UPC. The algorithm allows resynthesizing an input speech segment with the specifications of pitch and duration required.

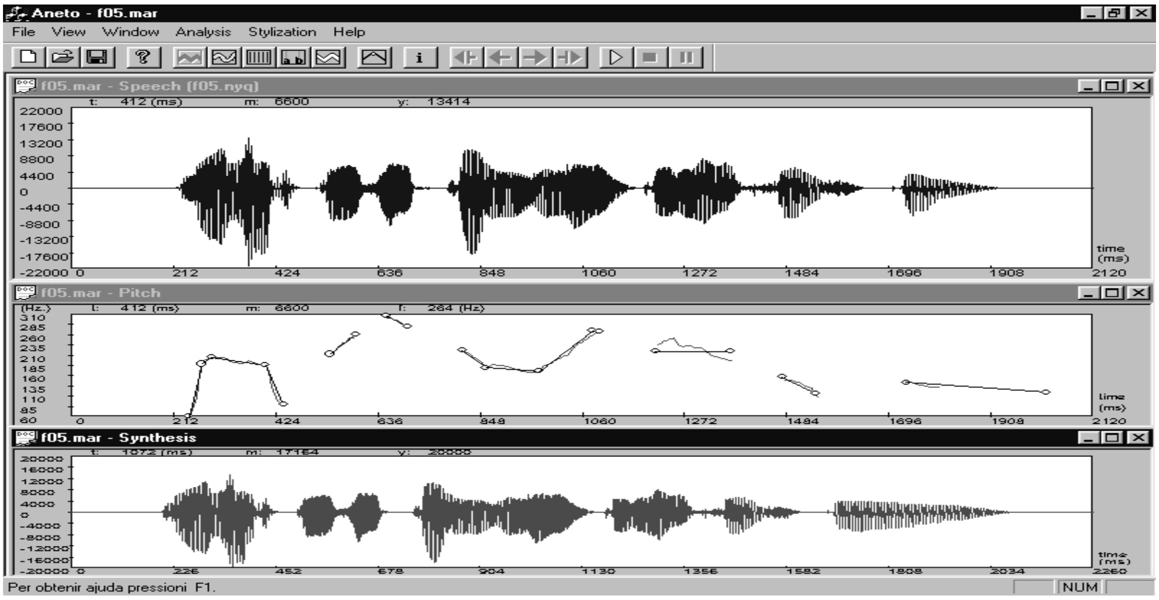

Figure 4. Analysis of a sentence with Aneto. First window corresponds to the speech file. Second window to pitch contour and stylization. Third window corresponds to resynthesized speech.

5 Labeling

Labeling is a common practice among people who study speech. It consists in putting some labels in order to store temporal information about the speech signal. An associated labeling window allows graphic edition of labels, usually marking the beginning or ending of phonetic segments. Marks can be inserted, moved, edited or deleted using the mouse and a pop-up menu. Labels can be personalized to different languages, prosodic notation systems or any other kind of label required by the user.

In the near future, automatic alignment will be added to the tool. Phonetic transcription can be derived from the text if the content of what has been said is known. Alignment based on Hidden Markov Models (HMM) will allow speeding the process of labeling. At the moment, it must be done manually.

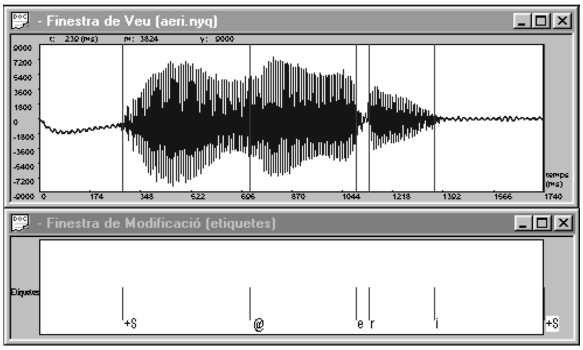

Figure 5. Phonetic labeling of a word (/@eri/).

6 Conclusions

Aneto is a software application that can analyze prosody of speech signals, stylize the pitch contour and label phonetic segments. After stylization, duration and fundamental frequency can be modified and by means of resynthesis with TD-PSOLA a speech signal with the new prosody can be obtained.

The tool has shown to be very useful in the study of prosody.

Future work will include automatic phone alignment, spectrogram representation and labeling of multiple speech signals of a database.

References

[1] A. Bonafonte, I. Esquerra, A. Febrer, F. Vallverdu, “A bilingual text-to-speech system in Spanish and Catalan”, *Proceedings of EuroSpeech'97*, pp. 2455-2458, Rhodes 1997.

[2] URL: http://gps-tsc.upc.es/veu/

[3] J.L. Navarro, I. Esquerra, “A Time-Frequency Approach to Epoch Detection”, Proceedings of Eurospeech'95, pp. 405-408, Madrid 1995.

[4] P. C. Bagshaw, *Automatic Prosodic Analysis for Computer Aided Pronunciation Teaching*, Ph.D., University of Edinburgh, 1994

[5] E. Moulines, F. Charpentier, “Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones”, *Speech communication*, 9, pp. 453-467, 1990.