AN ENVIRONMENT FOR THE LABELLING AND TESTING OF MELODIC ASPECTS OF SPEECH

Christel Brindöpke, Arno Pahde, Franz Kummert, Gerhard Sagerer
Technical Faculty, University of Bielefeld,
Postfach 10 01 31, 33501 Bielefeld
Germany
email: christel@techfak.uni-bielefeld.de

Abstract

In this paper, we present an environment for labelling and testing of melodic aspects of spoken language. The environment has three modes of application: First, the environment provides labelling facilities for a model-based melodic description for German. Second, it supports a language independent pre-theoretical description of speech melody allowing the development of new melodic categories. Third, our test bed can be used to generate speech samples with controlled melodic parameters for further use in perception experiments. The melodic description facilities (model-based, pre-theoretical) are supported by visual and audible feedback allowing a step-by-step refinement of the melodic description in question.

1 Introduction

Melodic aspects of speech are related with several linguistic and extralinguistic phenomenon. Linguistic aspects are for example the realization of accents or the marking of boundaries. Extralinguistic aspects concern e.g. speaker related qualities like emotions. Therefore the annotation of speech material with melodic items is useful and necessary for various applications in linguistic research and speech technology.

One of the notorical problems of melodic descriptions of speech material is the difficulty to test the fitting of the melodic description empirically. Our environment adresses to this problem by offering an audible feedback. However, an environment for melodic labelling which is conceptually based on visual and audible feedback has to take into account, that there is still no melodic model available that combines all factors that can influence speech melody. Therefore the environment needs modellbased definitions of melodic items on the one hand side and, on the other hand, has to provide a reasonable degree of freedom to allow modifications of the underlying melodic items or to enable the user to develop new melodic categories.

In the following section we outline the linguistic background of the three application modes of our environment. Section three and four describe our environment as such and section five shows first applications in all three modes. In the conclusions the results of the applications are discussed.

2 Linguistic background

For the mode of model-based labelling, our environment uses the pitch movements defined by Adriaens (1991) as predefined labels. This model is based on the work by 't Hart/Collier/Cohen (1990). They assume that in spoken language all perceptivly relevant changes in pitch can be described by means of a finite set of local and global pitch movements.

According to Adriaens (1991), the set of melodic items for German contains 12 local pitch movements and a declination component.The pitch movements are defined with respect to their position in the syllable (vowel onset, end of the voiced part of the syllable), their range (in semitones (ST)) and their duration (in milliseconds). Adriaens (1991) distinguishes five falling and seven rising pitch movements. As declination component Adriaens (1991) suggests five parallel declination lines in a logarithmic scale. The lines have a distance of 2.5 ST to each other. In non-final parts of utterances the overall decline of the lines is 6.1 ST and in final parts 8.5 ST. For the description of German speech melody the most relevant lines are the first, the third and the fourth declination line. The lines serve as a grid in which the local pitch movements are embedded.

It is known that the exact phonetic/acoustic definition of pitch movements depends also on factors like emotions or speaking rate. Shriberg et al. (1996)

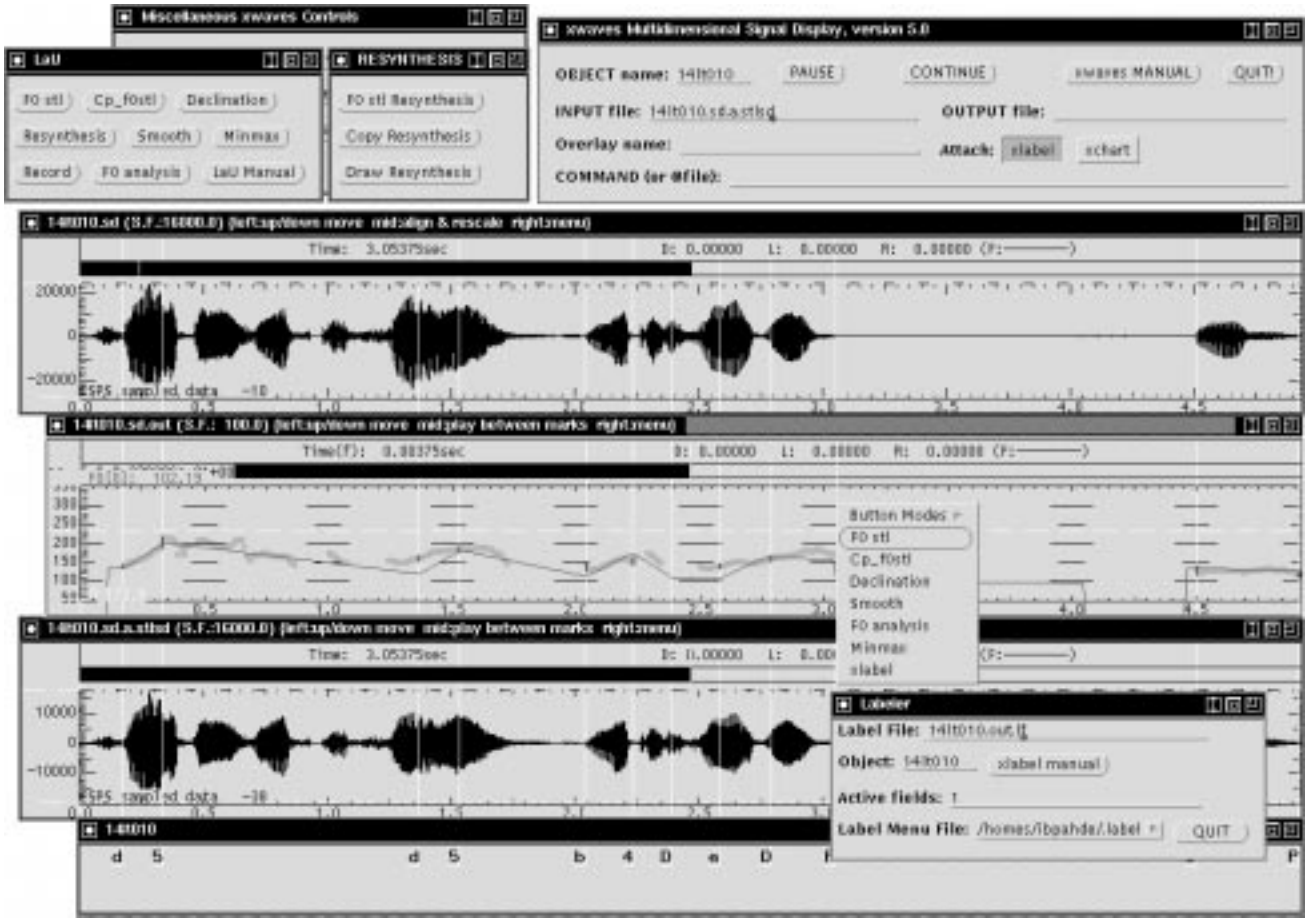

Figure 1 shows a screen dump of our test bed in the model-based labelling modus: The window in the second row displays the original speech signal, the window below shows the F0-contur of the speech signal and the calculated F0-contur (straight lines) as an overlay. The calculated F0-contur is derived from the choosen melodic labels given in the bottom display. The numbers denote different types of falling, the small letters denote different types of rising pitch movements, 'D' corresponds to a declination line. The window in the fourth row contains the speech signal, which has been resynthesised with the calculated F0-values. Further, the facilities offered by our environment can be seen in the menu-bars (mouse-menu, LaU- and resynthesis window).

have shown such an effect in the case of "Speaking Up" and Vroomen et al. (1993) investigate the relation between emotions and melodical and durational features of speech. Emotionally or fastly spoken utterances can have melodic attributes to the effect, that the helpfulness of the auditive feedback in the model-based labelling mode decreases. In the worst case it won't support the decision which melodic label to choose anymore. As a solution which allows the model-based melodic labelling of emotional or fast speech our environment offers in the mode of model-based labelling to insert scaling factors.

The pre-theoretical melodic description that our environment offers relies on the method of approximating the original F0-contour with a minimum set of straight lines. The aim here is to preserve the perceptual equality between the utterance with the original F0-contour and the utterance with the approximated F0-contour. A further analysis of an appropriate set of line segments can be used for the definition of new pitch movements (t' Hart/Collier/Cohen 1991).

For both kinds of the melodic description a resynthesis of the original speech utterance either with the model-based F0-values or with the F0-values which result from the pre-theoretical description is required. A well-known possibility to compute utterances with altered F0-contours is given by means of the Pitch Synchronous OverLAp technique developed by Moulines/Charpentier (1990).

3 The Labelling- and Testing Environment: LaU

The model-based labelling mode of our environment

offers a variety of choices which arise from the combination of the underlying melodic model and the degrees of freedom offered by the labelling environment. First of all, local and global pitch movement are computed seperately. That means, in our implementation, the user can make a melodical description in which local and global pitch movements follows Adriaens (1991) proposals (Figure 1) - in this case the environment offers a set of predefined labels - or he is free to combine the pre-defined local pitch movements with a declination component of his own choice.

To allow model-based melodic labelling even of very different speaking styles (emotional, very fast) our environment offers a scaling facility which allows an adapting of the underlying melodic items to the different speech styles.

The pre-theoretic description allows a melodic description in terms of user defined lines. The rise or fall of the line segments can be determined in frequency values or in semitones. This application mode can be used for the development of new melodic categories (e.g. for other languages than German) or it can be used to define any wanted F0-contour for further purposes.

Both kinds of melodic description (model-based, model-independant) are supported by a visual and an acoustic feedback mechanism (Figure 1). The aims of these feedbacks in the application mode of model-based labelling are twofold, firstly to facilitate the decision which label to choose and secondly, to make the process of melodic labelling intersubjective more reliable. In the mode of pre-theoretical melodic description the acoustic feedback can be used e.g. to test wether the aim of perceptual equality has been maintained or it can be used to compute speech signals with altered F0-contours [1]

4 Realization

The environment for the labelling and testing of melodic aspects of speech (Figure 2) is implemented in C and is integrated into the ESPS/XWaves environment. We have choosen to integrate our environment into ESPS/XWaves, because ESPS/XWaves is a world wide used speech analysis program that is available for various types of work-stations and that offers lots of processing and labelling facilities which now can be used in combination with our environment. For the needed resynthesis-facilities of LaU the PSOLA-algorithm [2] is used in the frequency domain.

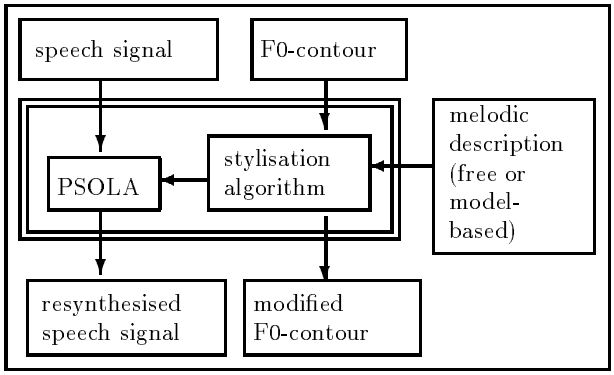

Figure 2 shows the main components of the labelling and test environment (surrounded by double lines) and its input and output files.

5 Applications

We tested our environment in its three domains of application: model-based melodic labelling, model-independant melodic description and the generation of speech samples with controlled melodic parameters. In the different application procedures two sets of German spontaneous utterances were used: the first corpus contains instruction dialogues in a man-machine-interaction (Brindöpke et al. 1995) and the second corpus consists of spontaneous instruction dialogues between two human dialog partners (Sagerer et al. 1994).

First, for the testing of the model-based labelling mode of our environment a set of 100 spontaneous German utterances which was selected of corpus one and two was labelled.

Second, the environment was tested in its application mode of pre-theoretical melodic description. For that purpose the F0-contours of a set of 30 spontaneous German utterances (samples out of the second corpus) were approximated in straight lines with the aim of preserving the perceptual equality.

For the testing of the environment as a tool for generating speech samples with controlled melodic parameters we supplied 48 resynthesised spontaneous utterances as test stimuli for a perception experiment. The aim of the perception experiment was to evaluate the melodic model which is used in our

[1] For more detailed description of the usage of the environment see Brindöpke/Pahde (1997). Roughly, the usage of the environment requires first to create files with the ESPS/XWaves-option 'xlabel' in which the choosen melodical description is written. The extensions of the files show which information they contain (e.g. '.lf' for files with model-based melodic labels, '.dci' for files with global pitch information). Then the options of the environment can be choosen in the mouse-menu (e.g. 'f0 stl' to compute the F0-contour which results from the melodic description, 'f0 stl resynthesis' to resynthesise the utterance with the new F0-contour).

[2] based on an PSOLA-implementation of Dik Hermes

environment for the model-based labelling mode for spontanous German speech. 24 spontaneous German utterances of the second corpus were resynthesised in two versions. In version one the speech melody was altered with the model-based labelling mode according to Adriaens (1991). In version two the local pitch movements were changed with the facilities of our pre-theoretical melodic description mode into British English speech melody according to Willems (1988) while the decline of the first declination line follows Adriaens (1991) proposal for German.

In neither of the three applications problems arise in the use of our environment. For final discussion of the application results see the following section.

6 Conclusions

Although the underlying model for the model-based labelling mode meets the demands of spontaneous German speech in general, two limitations can be observed: first, the positioning of some of the pitch movements was proven to show a larger degree of freedom than predicted by the model. This concerns especially those pitch movements whose position is defined with respect to the end of the voiced part of the syllable. Because the positioning of the melodic items is left to the responsibility of the user this weakness of the model does not concern the model-based labelling facilities of our environment. Second, the model does not account for the influences of emotions or speaking rate to the definition of the pitch movements. As it is defined, it meets the demands of spontaneous, not highly emotional German speech fairly well at a speaking rate of about 5.5 syllables per second. With increasing speech rate or in higly emotional speech the performance of the model decreases. This concerns especially the auditive feedback facilities of our environment: in the worst case, the auditive feedback can no longer be used for further improvement of the melodic labelling in question. Therefore for the model-based labelleling mode of our environment a scaling facility has been incorporated. The exact definition of the scaling factor is left to the user and can vary accordingly to the special characteristics of different speech data (highly emotional, very fast speech).

The second and third mode of application (pretheoretical melodic description, generation of speech samples with altered F0-contours) proved to be rather unproblematic. Desirable extensions of the labelling- and testing environment are for example the integration of a mechanism to control the prosodic parameter duration (e.g. PSOLA in the time domain).

References

[1] Adriaens, L.M.H. (1991). *Ein Modell deutscher Intonation. Eine experimentell-phonetische Untersuchung nach den perzeptiv relevanten Grundfrequenzveränderungen in vorgelesenem Text.* Dissertation, Technische Universität Eindhoven.

[2] Brindöpke, C./Johanntokrax, M./Pahde, A./Wrede, B. (1995).*Darf ich Dich Marvin nennen? Instruktionsdialoge in einem Wizard-of-Oz Szenario: Materialband.* Report 95/7, SFB 360: "Situierte Künstliche Kommunikatoren", Universität Bielefeld.

[3] Brindöpke, C./Pahde, A. (1997). *LaU: Label- und Testumgebung für melodische Aspekte gesprochener Sprache, Version 1.0.* Report 97/2. SFB 360: "Situierte Künstliche Kommunikatoren", Universität Bielefeld.

[4] Moulines, E./Charpentier, F. (1990). Pitch synchronous waveform processing techniques for text-to-speech synthesis using diphones. In: *Speech Communication 9 (1990)*, pp.453-467.

[5] Sagerer, G./Eikmeyer, H.J./Rickheit, G. (1994). *Wir bauen jetzt ein Flugzeug: Konstruieren im Dialog. Arbeitsmaterialien.* Tech. Rep., SFB 360: "Situierte Künstliche Kommunikatoren", Universität Bielefeld.

[6] Shriberg, E./Ladd, D.R./Terken, J./Stolcke, A.(1996). Modeling Pitch Range Variation within and across Speakers: Predicting F0 Targets when "Speaking Up". In: *Proceedings of the fourth International Conference on Speech and Language Processing*, Philadelphia 1996 pp. 1-4.

[7] 't Hart, J./Collier, R./Cohen, A. (1991). *A perceptual study of intonation. An experimental-phonetic approach to speech melody.* University Press, Cambridge.

[8] Vroomen, J./Collier, R./Mozziconacci, S. (1993). Duration and Intonation in Emotional Speech. In: *Proceedings of the third European Conference on Speech Communication and Technology,* Berlin 1993, pp. 577-580.

[9] Willems,N./Collier, R./'t Hart, J. (1988). A synthesis scheme for British English intonation. In: *Journal of the Acoustical Society of America, 84 (4), October 1988,* p. 1250-61.