AN ANNOTATION SYSTEM FOR MELODIC ASPECTS OF GERMAN SPONTANEOUS SPEECH

Christel Brindöpke, Brigitte Schaffranietz
Technical Faculty & Faculty of Linguistics and Literature
University of Bielefeld
christel@techfak.uni-bielefeld.de, brigitte@lili.uni-bielefeld.de

Abstract

This article presents a phonetically defined annotation system for German speech melody, whose descriptive units cover the perceptive relevant pitch movements. The units for the melodic annotation are based on the model for read German utterances [1]. The implementation of our descriptive melodic units as part of a recently developed labelling and testing environment for melodic aspects of speech allows a comfortable and intersubjective application of the melodic units for the annotation of speech. An experimental evaluation by a rating-experiment secures that the melodic units describe spontaneous speech as adequately as read speech

1 Introduction

During the last years there have been an increasing interest in the field of prosodic annotation of speech [8, 10, 12, 7, 6]. This originates from developments in speech technology and linguistic resarch of spontaneous speech.

We propose an annotation system expecially for melodic aspects of German speech whose units are perceptively relevant and phonetically defined. The process of defining the units from the speech signal and its calculated fundamental freuquency is determined by several selection criteria which secure the perceptive relevance of the units [13, 1]. To allow the application of the melodic units as an annotation system for spontaneous German utterances an experimental evaluation has been carried out to confirm the validity of the model for spontaneous German speech (section 3). Additionally, special labelling facilities for the annotation procedure are necessary. They have been developed, implemented in C and integrated in ESPS/XWaves. They allow a step-by-step refinement of the melodic description in question guided by visual and audible feedback (section 4). Because of the audible feedback by a resynthesis of the original speech signal with the model-based melodic units our annotation system is especially appropriate for testing hypotheses regarding the relation between speech melody and other linguistic levels of description (eg. syntax) and for descriptions of speech melody in the area of automatic speech processing.

2 Descriptive Units

As the perceptive relevance of the melodic units is a crucial point, a short outline of the defining procedure is given. One basic assumption is that in spoken language all perceptively relevant changes in pitch can be described by means of a finite set of local and global pitch movements [13]. The definition of those units from the speech signal and its calculated fundamental frequency is determined by three perceptive criteria which are applied at different levels of model building: *perceptual equality*, *perceptual equivalence* and *acceptability*. First, *close-copy-stylizations* of original F0-contours are made. A close-copy-stylization results from substituting the course of the original F0-contoru by a minimal number of straight lines. It has to be perceptively equal (auditive equal) to the original contour. The comparison of numerous close-copies leads to standard specifications of the perceptively relevant pitch movements. These *standardizations* have to be perceptively equivalent to the close-copies and the original contours. Whether the standardizations are *acceptable* melodic contours of the language under investigation and whether the close-copies are *perceptively equal* to the original contours is experimentally verified [9].

According to the outlined method [13, 1] for German a set of 14 descriptive units is proposed (see table1). The set contains 12 local pitch movements which are defined with respect to their position in the syllable, their range and their duration. 'D' represents contour segments following the global course (e.g. overall decline) of the F0-contour. 'P' represents intervals without relevance for the melodic description (silence, noise etc.). The combination rules of the descriptive units

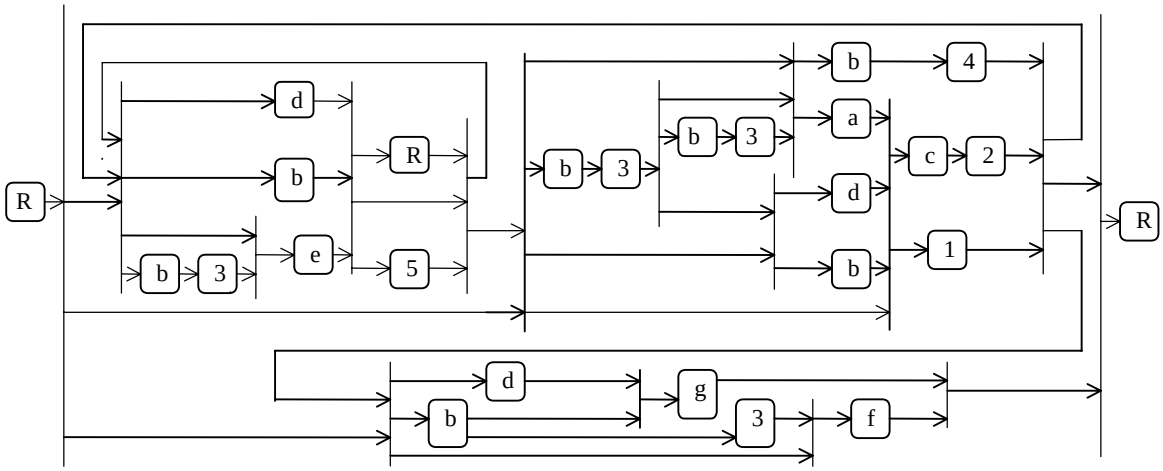

Figure 1: Network representing the linear order of the local pitch movements, R=reset.

can be described by a grammar and are represented in form of a transition network (figure 1). Speech segments following the declination line may occur at any position of the melodic contour. For convenience of clearness they are not integrated in the transition network. For more details see [5].

Label	Duration (ms)	Position (ms)	Size (st)	Func
a	180	vo-210	+7.5	acc
b	180	vo-60	+7.5	acc
c	60	vo-30	+2.5	acc
d	180	vo	+7.5	acc
e	180	evp-180	+7.5	bou
f	300	evp-300	+12.5	bou
g	120	evp-120	+5	bou
1	180	vo	-7.5	acc
2	240	vo+60	-10	acc
3	var	vo+120	-7.5	b_acc
4	180	vo+150	-7.5	acc
5	var	evp	-7.5	b_acc
D	var	var	decl	b_lpm
P	var	var	-	pause

Table 1: Pitch movements for German: vo=vowel onset, evp=end of voiced part of the syllable, st =semitones, var=variable, decl=overall decline of F0, ms=milliseconds, acc=accents, bou=boundary, b_acc=between accents, b_lpm=between local pitch movements.

3 Experimental Evaluation

In a rating experiment the validity of the melodic units for spontaneous German utterances was measured by the degree of the similarity between model-based and original contours. Similar experimental designs are described by [9, 13].

3.1 Experimental Design

In the experiment 24 native speakers were asked to judge the similarity of pairs of stimuli on a scale from 1 (unequal) to 7 (equal). The pairs of stimuli belong to either of three different conditions of similarity: group A contains pairs of identical contours, group B contains pairs of original and stylized German contours, group C contains pairs of either an original German and an English contour or a stylized German and an English contour.

The main point of interest is where the judgements of group B will be located on the scale between identical (group A) and not identical contours (group C). Provided that the model is valid for spontaneous German utterances, we would expect that the difference between the judgements for group A and group B should be significantly smaller than the difference between group B and C.

3.2 Stimuli, subjects and procedure

Out of 22 spontaneous instruction dialogues [11] 12 male and 12 female utterances were chosen according to the criteria different *syntactic structures* (simple declaratives, interrogatives, complex declaratives), *duration* between two and seven seconds and *occurrence* of all pitch movements of the model. For each condition of similarity (group A, B, C) 12 pairs of stimuli were built: every original utterance was resynthesized with the original F0 contour, the model-based F0 con-

tour and an English F0 contour according to [14]. The English contours should be as close to the German original as allowed by the English intonation model. The stylized German contours and the English contours were provided with an identical declination component in order to exclude possible influences caused by different declination lines. The melodic labelling and the resynthesis were done with an ESPS-based labelling and testing environment for melodic aspects of speech described in section 4.

3.3 Experimental results

For every subject decreasing rating values from condition A to B to C can be observed. As expected, the judgements of the conditions A, B, C differ significantly (for male stimul A to B: $t_{(23)}$=8.1410, B to C: $t_{(23)}$=14.677, p=0.000; for female stimuli A to B: $t_{(23)}$=8.583, p=0.000, B to C; $t_{(23)}$=16,617, p=0.000).

masculine voice			feminine voice		
cond.	mean	std.dev.	mean	std.dev.	n
A	6.45	.383	6.46	.566	24
B	5.44	.721	5.01	.822	24
C	2.52	1.074	2.65	.892	24

Table 2: Mean values for masculine and feminine stimuli.

Our main point of interest concentrates on the distance of A to B (A-B) and B to C (B-C) for male and female stimuli. ANOVA (with repeated measures, SPSS 7.5) shows a significance of the factors *sex* and *difference* with a main effect for *difference* ($F_{(1,23)}$=59.564, p=0.000). The distance of group A to B (A-B) is significantly smaller than the distance of group B to C (B-C): paired t-tests (2-tailed) for the factor *difference* show that A-B is significantly smaller than B-C for both sexes (male A-B/B-C: $t_{(23)}$=-8.435 , p=.000; female A-B/B-C: $t_{(23)}$=-3.915, p=.001). These results give sufficient support for the validity of the model for spontaneous spoken German utterances.

4 Application of the melodic annotation system

To allow the melodic annotation with the pitch movements proposed in section 2 appropriate labelling facilities must be available. Therefore, a labelling and testing environment for melodic aspects of speech has been developed [3, 2]. The melodic model has been implemented as part of the environment. This environment offers audible and visual feedback for the melodic annotation thus allowing a step-by-step refinement of the melodic description in question.

4.1 Audible and visual feedback

The audible feedback consists of a resynthesis of the original speech signal. This resynthesis has an altered fundamental frequency which is computed from the melodic annotation supplied by the user. By auditive comparison of the resynthezised speech signal with the original, it is possible to decide whether the melodic annotation is an appropriate description of the original speech signal.

The visual feedback shows the fundamental frequency computed from the melodic annotation as an overlay on the original fundamental frequency and allows a visual comparison of the original and the model-based contour as well. For a successful comparison with respect to the similarity to the original and the computed fundamental frequency it should be taken into account that not all visual differences between two contours are audible as well.

4.2 Scaling factors and other degrees of freedom

Generalizing over the melodic variety in speech is an important advantage of a model-based melodic description. However, since already the model building itself is based on a restricted set of data such a general description can miss the fine grained characteristics of special corpora (sociolects, ideolects or emphasized speech). To describe such fine grained melodic deviations without loosing the generalizing abilities of a melodic model our labelling facilities provide scaling factors and other degrees of freedom. The user can adapt the local pitch movements by inserting scaling factors. Because of its varying pitch range this is an example for the usefulness and necessity of the scaling of the local pitch movements. For computing the fundamental frequency which results from the melodic description a declination model following [1] was implemented. Since the slope of declination is known to vary according to several variables like position in a paragraph, utterance length etc. the user can define his own declination line by giving either its starting value and the overall decline (in semi tones) or starting and end value.

4.3 Realization and Usage

The environment for the labelling and testing of melodic aspects of speech (see figure 2) is implemented in C and integrated into the ESPS/XWaves environment. ESPS/XWaves is a world wide used speech analysis program that is available for various types of work-stations. It offers lots of processing and labelling facilities which can be used in combination with our

environment. The resynthesis-facilities of our environment include the PSOLA-algorithm in the frequency domain implemented by Dik Hermes.

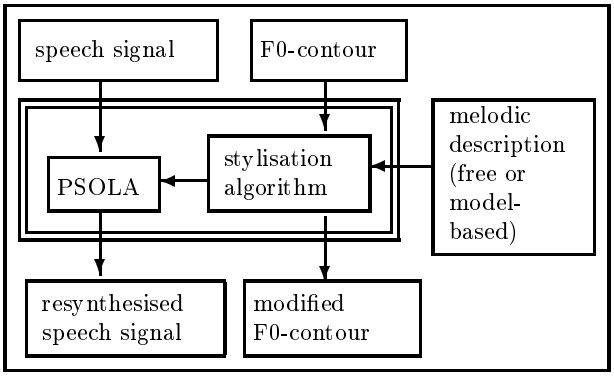

Figure 2: Main components of the labelling and testing environment (surrounded by double lines) and its input and output files.

Roughly, the usage of the environment requires files which contain the melodic annotation of the speech signal with the pitch movements given in table 1. This annotation is provided by the user according to his auditive and visual analysis of the speech signal. The files are created by using the ESPS/XWaves facilities for labelling (e.g. 'xlabel'). One obligatory file contains the melodic annotation of the speech signal and a second optional file contains information about the global course of the fundamental frequency. The extensions of the files indicate which information they contain (e.g. '.lf' for files with melodic labels, '.dci' for files with global pitch information). The options of the label environment e.g. the auditive or the visual feedback are chosen and executed via mouse menue. For a more detailed description of the usage and additional options provided by this labelling environment see [2].

5 Conclusion

We have presented an annotation system for German speech melody which is methodically based on the IPO-approach. The implementation of the model as part of a labelling and testing environment and the experimental evaluation allow the broad application of the perceptively relevant and phonetically defined descriptive units as an annotation system for German speech melody. The labelling and testing environment supports an intersubjective use of the pitch movements. Aims of further research concern the expansion of the currently defined pitch movements as well as the improvement of the declination component which is currently used in the implementation of the labelling environment.

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./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.

[3] Brindöpke, C./Pahde, A./Kummert, F./Sagerer, G. (1997). An environment for the labelling and testing of melodic aspects of speech. *Proceedings of the fifth European Conference on Speech Communicatication*, Rhodes, pp. 199-202.

[4] Brindöpke, C./Schaffranietz, B. (1997). Evaluation of an intonation model for German spontaneous speech. *Proceedings ESCA workshop on intonation*, Athens, Greece Sept. 18-20, pp. 51-54.

[5] Brindöpke, C./Schaffranietz, B. Ein Transkriptionssystem für die Sprachmelodie des Deutschen. Linguistische Berichte. (forthcoming).

[6] Grice, M./Benzmüller, R. (1995). Transcription of German using ToBI tones: The Saarbrücken system. PHONUS, Institut für Saarbrücken.

[7] Kohler, K. (1995). PROLAB - the Kiel system of prosodic labelling. *Proceedings of International Congress of Phonetic Sciences*, Stockholm, vo l.3 pp. 162-165.

[8] Mayer, J.(1995). *Transcription of German intonation: The Stuttgart system.* Technischer Bericht. Institut für maschinelle Sprachverarbeitung Universität Stuttgart.

[9] de Pijper, J.R. (1983). Modelling British English Intonation. An analysis by resynthesis of British English Intonation. Dordrecht/ Cinnaminson: Foris.

[10] Reyelt, M./Grice, M./Benzmüller, R./Mayer, J./Batliner, A. (1996) Prosodische Ettikettierung mit GToBI. Gibbon, D. (ed.) (1996), *Natural Language Processing and Speech Technology. Results of the third KONVENS conference.* Berlin: de Gruyter, pp. 144-155

[11] 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.

[12] Selting, M. (1995) *Prosodie im Gespräch. Aspekte einer interaktionalen Phonologie der Konversation.* Tübingen: Niemeyer.

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

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