A HMM-BASED RECOGNITION SYSTEM FOR PERCEPTIVE RELEVANT PITCH MOVEMENTS OF SPONTANEOUS GERMAN SPEECH[*]

Christel Brindöpke, Gernot A. Fink, Franz Kummert, Gerhard Sagerer
Technical Faculty, University of Bielefeld
christel@techfak.uni-bielefeld.de

Abstract

This paper presents an HMM-based recognition system for perceptive relevant pitch movements of spontaneous German speech. The pitch movements are defined according to the perceptively and phonetically motivated IPO-approach to intonation. For recognition we use a hybrid approach combining polynomial classification with Hidden Markov Modelling. The recognition is based only on the speech signal, its fundamental frequency and eleven derived features. We evaluate the system on a speaker independant recognition task.

1 Introduction

In current speech recognition systems, usually, no prosodic information is used. However, it is a well-known assumption that prosody can contribute useful information to enhance speech recognition and understanding processes [6]. While for speech recognition, there is no doubt that recognition processes should be resulting in a sequence of word hypotheses for prosody recognition the chosen units depend on several competing linguistic theories. Although their advantages and disadvantages on a level of linguistic description have been discussed thoroughly, further exploration of their effectiveness in the area of automatic speech processing is necessary.

So far most approaches on German prosody recognition have concentrated on the recognition of accents and phrase boundaries [6, 9, 8]. In contrast to that our aim is the recognition of melodic aspects of speech. The aim is to provide a melodic description which covers the complete speech signal in terms of melodic items. Since for languages like English or German speech melody is strongly connected to accent and boundaries our approach can provide information about these kinds of segments by defining the relation between speech melody and accents and boundaries. For the automatic recognition of speech melody, we propose intonational units that are defined according to the phonetically and perceptively motivated IPO-approach to intonation [10]. The resulting pitch movements describe the perceptive relevant melodic aspects of German speech. Their validity was tested in perception experiments for read-aloud speech [1] and later we have shown in perception experiments that they are valid for spontaneous speech as well [3]. Section 2 gives an overview of the underlying melodic model. For the recognition we use a hybrid system combining polynominal classification with Hidden Markov Models (HMMs) which is described in section 3. Section 4 describes the speech corpus currently used for training and testing purposes and experimental results are given in section 5.

2 Underlying melodic model

As the perceptive relevance of the melodic units is a crucial point, a short outline of the defining procedure is given.

Based on the assumption 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 the process of defining those pitch movements consists of a step-by-step reduction of measured F0-contours. This process is guided by three perceptive criteria *perceptual equality*, *perceptual equivalence* and *acceptability* which are applied at different levels of model building and shall ensure the perceptive relevance of the melodic items gained in the reduction process. Wether these three criteria are met can be and has been verified experimentally. Proposals and discussions of several experimental setups are given in [4, 10].

[*] This research was partly supported by a graduate grant of the Ministry of Science and Research (MWF) of the state North-Rhine-Westfalia and by the German Research Foundation (DFG) within the SFB 360 at the University Bielefeld. Only the authors are responsible for the contents of this publication.

Firstly, in the reduction process the course of original F0-contours (logarithmic scale) is substituted by a minimal number of straight lines. The resulting contours have to be perceptively equal to the original contours. Secondly, the comparison of numerous of those "substituted" contours leads to proposals for standard spezifications of perceptively relevant pitch movements of the language under investigation. Contours made of those standardizations have to be perceptively equivalent to the original contours or to those by straight lines approximated contours. Additionally, these standardized melodic contours have to be acceptable contours of the language under investigation.

For German, a set of 14 descriptive units is proposed (see table 1). 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 the pause which means in this case silence, background noise etc. The syntagmatic relation of the melodic units can be described by means of a regular or context free grammar.

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

Table 1: Pitch movements for German: vowel onset (vo), end of voiced part of the syllable (evp), semitones (st), variable (var), overall decline of F0 (decl), milliseconds (ms).

function	labels
accent	a, b, c, d, 1, 2, 4
boundary	e, f, g (3, 5)
between accents or boundaries	3, 5, D

Table 2: Functions of pitch movements.

As mentioned above there is a strong connection between speech melody and prosodic items like accents or the boundary signalling pitch movements. Every pitch movement can be interpreted as an accent lending pitch movement, a boundary signalling pitch movement or as a pitch movement connecting occurences of the two types. As it is shown in table 2, the pitch movements of type 3 and 5 occur as a boundary signal as well as they connect pitch movements of other types. Thus, the pitch movements on their own contribute important information to further usage in speech recognition and understanding systems. Besides, the overall melodic pattern which they form is strongly connected to aspects of discourse structuring, syntax and semantics [10, 11].

3 Outline of the recognizer

The recognition system consists of feature extraction, polynomial classification and HMM based recognition. It uses the speech signal (raw data, 16000 Hz sampling frequency) as its input and provides one melodic description per speech signal (best chain).

3.1 Feature extraction

Firstly, for feature extraction the fundamental frequency is calculated using the integrated pitch algorithm of ESPS/XWaves. For the purpose of recognition we use 11 dimensional feature vectors as proposed by [9][1]. They are calculated with a frame rate of 10 ms and each of them describes fundamental frequency and energy contours of the speech signal over an interval of 200 ms. The fundamental frequency is interpolated and decomposed into three components by band pass filters. Time derivatives of the interpolated F0 contour and its three components describe the F0 contour locally and globally. To this eight features per frame which are based on fundamental frequency, 3 energy features, the so called nasal band, sonorant band and the fricative band are added.

3.2 Classification

For the calculation of the emission probabilities of the HMMs we use a hybrid approach combining a polynomial classifier with Hidden Markov Models. Usually, the goal of a classification system is the mapping of a feature vector $\vec{c}$ into one of K classes Ω_k. The polynomial classifier approximates the perfect classification functions

$$\delta_k(\vec{c}) = \begin{cases} 1 & \text{if} \quad \vec{c} \in \Omega_k \\ 0 & \text{otherwise} \end{cases}$$

[1] We wish to express our thanks to Volker Strom, IKP, University Bonn, who kindly provided the software for feature extraction.

by a polynomial in the coefficents of the feature vector. Let $\vec{x}(\vec{c}) = (1, c_1, \ldots, c_n, c_1^2, c_1c_2, \ldots, c_n^2, c_1^3, \ldots)^T$ denote the expanded feature vector then the estimated classification functions can be written as

$$d_k(\vec{c}) = \vec{a}_k^T \vec{x}(\vec{c}).$$

The optimal parameter vectors $\vec{a}_k$ are calculated on the basis of a classified training sample by minimizing the mean square error between the perfect and the estimated classification functions. According to the Weierstrass Theorem, arbitrary functions can be approximated in such a way where the accuracy only depends on the degree of the polynomial.

3.3 Hidden-Markov Modelling

For the modelling of prosodic events on a segmental level we use HMMs. Basic methods for model building, parameter training and decoding are provided by a general HMM toolkit which we normally use for the design of our speech recognition systems [5, 12]. This framework in general supports continuous mixture models that can be structured hierarchically and can use either data or model-driven strategies for parameter tying. During Viterbi-decoding statistical language models or context free grammars can be applied to constrain the search process.

As currently we do not yet have sufficient knowledge about a possible internal structure of the prosodic events described in table 1, for each of them a single basic model with a certain number of tied states is initialized from the labelled data. In order to account for three phases - transitions to left and right context and center, as commonly used for phonetic units - and for widely differing segment durations up to four corresponding models are built for each segment as concatenations of a certain number of the appropriate basic units. Initially, all of them have identical parameters. However, during Baum-Welch reestimation parameters of compound models are adjusted independently.

Our hybrid approach is similar to a semi-continous HMM (SCHMM) where instead of a mixture of gaussian distributions a mixture of estimated classification functions $d_k(\vec{c})$ is used. The mixture coefficients $b_{ik}, i = 1..N, k = 1..K$ (N number of HMM states, K number of polynomials) are estimated during the Baum-Welch training analogously to SCHMMs. Therefore, the use of a polynomial classifier of second degree is in principle equivalent to a conventional SCHMM because the expression $(\vec{c} - \vec{\mu})^T \underline{K}^{-1} (\vec{c} - \vec{\mu})$ in a gaussian distribution is a polynomial of second degree in the coefficents of the feature vector $\vec{c}$.

4 Speech Corpus

The speech corpus we currently use for training and testing purposes is taken from a larger set of recorded man-machine interaction. In contrast to other approaches (e.g [7]) where the speech corpus was elicitated especially with the aim of providing optimal intonation patterns our corpus has been recorded in a Wizard-of-Oz scenario whose main aim was to provide spontaneous man-machine-dialogues within the domain of a construction scenario [2]. The task for the subjects was to advise the computer to build a toy airplane. During the dialogue the resulting intermediate construction stages had been visualized on the screen. Output of the simulated speech system had been predefined sentences realized via a speech synthesis system. The whole corpus consists of 50 recorded man-machine dialogues containing approximately 8 hours of speech (51330 words) spoken by 34 male and 16 female speakers.

Since the melodic annotation of speech signals is a very time consuming task, we currently use approximately 30 minutes of speech containing speech samples of 30 speakers for training and 5 minutes of speech by 10 different speakers for evaluation purposes. Since the corpus was not recorded especially with the aim of providing optimal intonation contours, the occurences of the melodic items vary significantly according to their acoustic realization. In addition, some types of pitch movements occur rather frequently whereas others occur only a few times. The frequency of occurences in the corpus currently used varies from a few to approximately 1000 occurences for the larger classes. However, this phenomenon is not specific to our training material but reflects the occurence frequency of the pitch movements in general.

5 Experiments

The set up for the experiments done so far, was the recognition task of all 14 classes given in table 1.

First experiments have been performed using unsupervised learning for codebook estimation. A codebook for mixture density classification was generated by an unsupervised clustering of 2,5 hours of spontaneous speech of our corpus using the LBG-algorithm to construct a set of 64 gaussian distributions with full covariance matrices. Using this codebook for a mixture density classification has not shown satisfactory results. Presumably this is due to widely differing class sizes of the pitch movements resulting in a prior probability close to zero for most of the estimated classes.

Therefore we choose the approach described in section 3 which seems to be more robust towards varying class sizes. The 11 dimensional feature vectors are classified into 14 classes corresponding to the 14 types of pitch movements (table 1) using a polynomial classifier of fifth degree. For HMM-modelling currently one state per basic model is initialized. For most of the pitch movements three phases are assumed and at least one model for short durations and one for longer durations is given. Currently, no grammar for pitch movements is used – a contour is defined as an arbitrary sequence of the 14 types of pitch movements. Table 1 gives an overview of the corpora used for training and testing purposes. Table 2 shows the recognition results calculated on the best chain. The system is evaluated as a conventional continuous speech recognizer using a 14 word lexicon. Word accuracy and some related measurements have been calculated.

			speaker	
corpus	turns	min	male	female
training	150	30	21	9
test	23	5	6	4

Table 1: Corpora for training and testing

	results				
corpus	wa	wc	sub	del	ins
training	54.92	55.62	10.55	33.83	0.70
test	48.28	49.75	19.21	31.03	1.48

Table 2: Recognition results for training and test, word accuracy (wa), words correct (wc), substitutions (sub), deletions (del), insertions (ins).

Though the recognition performance is still low the results show that the approach is feasable. Most of the errors are due to the class 'D'. This class can be considered as a superclass of a set of very heterogeneous subclasses which can not yet be defined explicitely. However, the correct recognition yields segment boundaries which are close to the segment boundaries of manual labelling. This makes it possible to use the outlined system despite its currently low performance as an initial guess for further manual labelling.

6 Conclusion

This paper presented a hybrid approach for the recognition of perceptively relevant pitch movements of German spontaneous speech. Currently, the design of the approach focuses more on the problems rised by spontaneous speech, speaker independancy and varying class sizes than on the question of optimizing recognition results. The results show that in general the approach is feasable. However, further improvement of the recognition performance is necessary. Especially, increasing the corpus of melodically annotated speech will enable us to investigate the use of a statistical language model for improving the recognition performance.

References

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

[2] C. Brindöpke, J. Häger, M. Johanntokrax, A. Pahde, M. Schwalbe, and B. Wrede. Darf ich dich Marvin nennen? Instruktionsdialoge in einem Wizard-of-Oz Szenario: Szenario-Design und Auswertung. Report 95/16, SFB 360: 'Situierte Künstliche Kommunikatoren', Universität Bielefeld, 1995.

[3] C. Brindöpke and B. Schaffranitz. Evaluation of an intonation model for german spontaneous speech. In *Proceedings ESCA Workshop on Intonation, September 18-20*, pages 51–55, Athen, 1997.

[4] J. R. de Pijper. *Modelling British English intonation. An analysis by resynthesis of British English intonation.* PhD thesis, Technische Universität Eindhoven, 1983.

[5] G. A. Fink, C. Schillo, F. Kummert, and G. Sagerer. Incremental speech recognition for multimodal interfaces. In *IECON*, Aachen, 1998. to appear.

[6] W. Hess, A. Batliner, A. Kiessling, R. Kompe, E. Nöth, A. Petzold, M. Reyelt, and V. Strom. Prosodic modules for speech recognition and understanding in VERBMOBIL. In Y. Sagisaka, N. Campbell, and N. Higuchi, editors, *Computational models for processing spontaneous speech*, pages 361–382. Springer, Berlin, 1997.

[7] U. Jensen, R. K. Moore, P. Dalsgaard, and B. Lindberg. Modelling of intonation contours at the sentence level using CHMMS and the 1961 o' Connor and Arnold scheme. In *Proceedings of the 3rd European Conference on Speech Communication and Technology*, pages 785–788, Berlin, 1993.

[8] R. Kompe, A. Batliner, A. Kießling, U. Kilian, H. Niemann, E. Nöth, and P. Regel-Brietzmann. Automatic classification of prosodically marked phrase boundaries in German. In *Proceedings Internatinal Conference on Acoustic, Speech and Signal Processing*, volume 2, pages 173–176, Adelaide, 1994.

[9] V. Strom. Detection of accents, phrase boundaries and sentence modality in German with prosodic features. In *Proceedings of the fourth European Conference on Speech Communication and Technology*, pages 2039–2041, Madrid, 1995.

[10] J. t' Hart, R. Collier, and A. Cohen. *A perceptual study of intonation. An experimental-phonetic approach to speech melody.* Cambridge University Press, Cambridge, 1990.

[11] J. Terken and R. Collier. Syntactic influences on prosody. In Y. Tohkura, E. Vatikiotis-Bateson, and Y. Sagisaka, editors, *Speech Perception, Production and Linguistic Structure*, pages 427–438. Ohmscha & IOS Press, Tokyo, Amsterdam, 1992.

[12] S. Wachsmuth, G. A. Fink, and G. Sagerer. Integration of parsing and incremental speech recognition. In *Proc. European Signal Processing Conference*, Rhodes, 1998. to appear.