A COMPARATIVE STUDY OF HMM-BASED APPROACHES FOR THE AUTOMATIC RECOGNITION OF PERCEPTUALLY RELEVANT ASPECTS OF SPONTANEOUS GERMAN SPEECH MELODY*

Christel Brindöpke, Gernot A. Fink, Franz Kummert
Technical Faculty, University of Bielefeld,
P.O. Box 100131, 33501 Bielefeld, Germany
christel@techfak.uni-bielefeld.de

ABSTRACT

Three approaches to the speaker independent automatic recognition of melodic aspects of spontaneous German are presented. All systems are based on Hidden Markov Models. Their input is restricted to the speech signal from which a feature extraction component derives eleven prosodic features. No additional information – as commonly used for prosody recognition – like word chains, word hypotheses, further segmental or lexical prosodic information (e.g. stress placement) is required. The three systems are tested and compared with respect to their performance on a speaker-independent recognition task on spontaneous German speech focusing on three functional aspects of speech melody (accent lending, boundary signalling, concatenating pitch movements) and the pause as a fourth category.

Our aim is to provide a melodic description which completely covers the speech signal in terms of melodic categories. Melodic patterns are strongly connected to aspects of discourse structuring, syntax and semantics [8, 9] and are of major importance in speech recognition and understanding processes. However, since not every single pitch movement has been explored yet with respect to its use and its meaning for speech recognition and understanding we reduce the melodic information to three main classes whose functions are already known. Therefore we distinguish between accent lending melodic units, boundary signalling and concatenating pitch movements. We add the pause as a fourth prosodic category in order to cover the whole speech signal and to offer the structuring information connected with this category for further speech recognition and understanding processes.

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 [5]. 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.

Section 2 gives an overview of the underlying melodic model and the relation between the defined melodic units and the three functional classes. Section 3 describes three different recognition approaches. We explore commonly used and well-known methods from the area of speech recognition. Currently, our systems do not apply additional knowledge which would allow to predict the prosodic categories mentioned above. This means we do not use knowledge of the word chain, the position of the word in the utterance or the grammatical category of the word since all this information would already allow to make predictions of accent lending or concatenating pitch movements in the speech signal. Section 4 describes the speech corpus currently used for training and testing purposes. Experimental results and a comparison of the systems are given in section 5.

*This research was partly supported by a graduate grant of the Ministry of Science and Research (MWF) of the state North-Rhine-Westfalia.
2 MELODIC MODEL

Based on the assumption that in spoken language all perceptually relevant changes in pitch can be described by means of a finite set of local and global pitch movements a process of defining those pitch movements was developed [8]. It consists of a step-by-step reduction of measured fundamental frequency contours (F0-Contours) guided by perceptive criteria to ensure the perceptive relevance of the melodic units.

For German, a set of 13 descriptive melodic units has been proposed [1] and evaluated for spontaneous speech [3]. The units are defined with respect to their position in the syllable, their range and their duration. Every pitch movement can be interpreted as an accent leading pitch movement, a boundary signalling pitch movement or as a pitch movement concatenating occurrences of the two types (table 1). The last type frequently occurs in stretches of speech containing un-accented syllables. In a semantic/pragmatic analysis, normally, these segments do not reveal additional or new information.

<table>
<thead>
<tr>
<th>categories</th>
<th>descriptive melodic units</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Duration (ms)</td>
</tr>
<tr>
<td>accent</td>
<td>180</td>
</tr>
<tr>
<td></td>
<td>180</td>
</tr>
<tr>
<td></td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>180</td>
</tr>
<tr>
<td></td>
<td>180</td>
</tr>
<tr>
<td></td>
<td>240</td>
</tr>
<tr>
<td></td>
<td>180</td>
</tr>
<tr>
<td>boundary</td>
<td>180</td>
</tr>
<tr>
<td></td>
<td>300</td>
</tr>
<tr>
<td></td>
<td>120</td>
</tr>
<tr>
<td>concatenating or</td>
<td>var</td>
</tr>
<tr>
<td>boundary</td>
<td>var</td>
</tr>
<tr>
<td>pause</td>
<td>var</td>
</tr>
</tbody>
</table>

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

The functional disambiguation of the last three melodic units depends on the melodic context. For example, if the first pitch movement in this functional category is followed by a pause, then it is highly probable that it will indicate a boundary otherwise it will indicate a concatenating function.

3 RECOGNITION SYSTEMS

The HMM-based recognition systems for the four prosodic categories share the same overall design consisting of a feature extraction stage and a classification stage. The common feature extraction stage and the model structure used are presented in the following two subsections. They are only differing with respect to the modelling of the state dependent emission probability densities. In section 3.3 and 3.4 respectively the approaches investigated will be described.

3.1 FEATURE EXTRACTION

For feature extraction the fundamental frequency is calculated using the integrated pitch algorithm of ESPS/XWaves. The recognition is based on 11-dimensional feature vectors as proposed by [7]. They are calculated with a frame rate of 10 ms and describe 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 HMM-MODELLING

For the modelling of our prosodic categories 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 [4, 10]. 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 initialised from the labelled data. In order to account for three phases - transitions to left and right context and centre, 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 re-estimation parameters of compound models are adjusted independently.

1We wish to express our thanks to Volker Strom, IKP, University of Bonn, who kindly provided the software for feature extraction.
3.3 SCHMM- AND CDHMM-BASED APPROACH

In most current HMM systems emission probability densities are modelled using mixtures of Gaussians. Depending on whether the Gaussians are specific to a single mixture density or are shared between all densities the corresponding models are called continuous mixture HMMs (CDHMM) or semi-continuous HMMs (SCHMM). The common set of Gaussians used in a SCHMM system is usually called the codebook. In the CDHMM case every state uses individual Gaussians from a local codebook.

In both modelling approaches the initial parameters of the codebooks are of fundamental importance and have to be estimated by applying a vector quantisation algorithm to appropriately labelled training data. For our initial codebooks we used the LBG-algorithm [6] to generate separate codebooks for each of the four categories to be recognised. In the SCHMM system all resulting Gaussians were merged into a single codebook. A completely unsupervised optimisation of a single codebook exhibited rather poor performance in informal tests, which is most likely due to limited training material for classes observed infrequently.

The remaining HMM parameters, namely the mixture weights and state transition probabilities are initialised on labelled data and subsequently optimised using Baum-Welch re-estimation.

3.4 HYBRID APPROACH

Our hybrid approach is similar to a semi-continuous HMM (SCHMM) where instead of a mixture of Gaussian distributions a mixture of estimated classification functions \( d_k(\hat{c}) \) is used. The mixture coefficients \( b_{k,i} = 1, N, k = 1..K \) (\( N \) 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 \( (\hat{c} - \mu)^T K^{-1} (\hat{c} - \mu) \) in a Gaussian distribution is a polynomial of second degree in the coefficients of the feature vector \( \hat{c} \).

Usually, the goal of a classification system is the mapping of a feature vector \( \hat{c} \) into one of \( K \) classes \( \Omega_k \). The polynomial classifier approximates the perfect classification functions

\[
\delta_k(\hat{c}) = \begin{cases} 
1 & \text{if } \hat{c} \in \Omega_k \\
0 & \text{otherwise}
\end{cases}
\]

by a polynomial in the coefficients of the feature vector. Let

\[
\hat{x}(\hat{c}) = (1,c_1,\ldots,c_n,c_1^2,\ldots,c_n^2,c_1^3,\ldots)^T
\]

denote the expanded feature vector then the estimated classification functions can be written as

\[
d_k(\hat{c}) = a_k^T \hat{x}(\hat{c}).
\]

The optimal parameter vectors \( a_k \) are calculated on the basis of a classified training sample by minimising 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.

4 SPEECH CORPUS

The speech corpus for training and testing purposes is taken from a larger set of recorded man-machine interaction. It has been recorded in a Wizard-of-Oz scenario whose main aim was to provide spontaneous task-oriented man-machine dialogues within the domain of an assembly scenario [2]. The task for the subjects was to advise the computer to build a toy airplane. During the dialogue the resulting intermediate assembly stages were visualised on the screen. Output of the simulated speech system were pre-defined sentences realized by 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 45 minutes of speech containing speech samples of 30 speakers for training and 8 minutes of speech by 10 different speakers for evaluation purposes. The training and testing material has been selected randomly from the larger corpus. The occurrences of the melodic items vary significantly according to their acoustic realization. The training corpus contains only approximately two hundred boundary signalling pitch movements and up to ten times more training samples for the other three classes.

5 EXPERIMENTS

The three systems were tested on a four class recognition task. They classify the 11-dimensional feature vectors into one of four categories: accent-lending, boundary-signalling, concatenating pitch movements, or the pause. The HMM-modelling, we currently use, is fairly simple: 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. A contour is defined as an arbitrary sequence of the three melodic categories and the pause. Table 2 gives an overview of the corpora used for training and testing purposes.
Table 2: Speech material for training and testing

<table>
<thead>
<tr>
<th>corpus</th>
<th>turns</th>
<th>min</th>
<th>speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td>training</td>
<td>240</td>
<td>45</td>
<td>male</td>
</tr>
<tr>
<td>test</td>
<td>39</td>
<td>8</td>
<td>6</td>
</tr>
</tbody>
</table>

Recognition results for the three approaches calculated on the best chain are shown in table 3. The systems are evaluated like a conventional continuous speech recogniser using a 4 word lexicon.

<table>
<thead>
<tr>
<th>system</th>
<th>words correct</th>
<th>word accuracy</th>
<th>RTF</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCHMM</td>
<td>66.25</td>
<td>58.28</td>
<td>0.16</td>
</tr>
<tr>
<td>CDHMM</td>
<td>77.34</td>
<td>66.41</td>
<td>0.15</td>
</tr>
<tr>
<td>Hybrid</td>
<td>80.47</td>
<td>70.00</td>
<td>0.45</td>
</tr>
</tbody>
</table>

Table 3: Recognition results of the three systems on the independent test set and the real time factors (RTF) required for processing.

The SCHMM and CDHMM systems use a total of 283 Gaussians with diagonal covariance matrices either as a single shared codebook or as four separate codebooks for the four prosodic categories. In the hybrid system the Gaussians are replaced by a polynomial classifier of 5th degree that approximates the discrimination functions between the four categories. The degree of the classifier was determined in informal experiments showing that with 6th degree already significant over-adaptation occurs.

Though the SCHMM and CDHMM systems use the same initial Gaussians, the CDHMM system performs significantly better. This is mainly due to the fact that codebooks for different events are not merged for emission probability calculations which can be considered as a kind of smoothing operation.

The best results are achieved using the hybrid approach which strongly suggests that even a mixture of a limited amount of Gaussians does not reflect the regions of prosodic events in feature space accurately enough. However, in contrast to estimating Gaussian mixture models the polynomial classification approach requires extremely demanding computations during the training. In order to obtain the optimal parameters for a 5th degree classifier using 11 features the calculation and inversion of a $4372 \times 4372$ moment-matrix is required. Therefore the hybrid approach does not easily scale up with an extended feature set. During the recognition process the efficiency of the SCHMM-, the CDHMM-, and the hybrid system are comparable. All three recognisers run faster than real time on a DIGITAL Alpha station 500/433.

6 CONCLUSION

We presented HMM-based approaches for the recognition of melodic aspects of spontaneous German speech. For the modelling of emission probability densities, Gaussian mixture models within traditional HMM systems were compared to a hybrid approach using a polynomial classifier. In a speaker independent recognition experiment using 4 prosodic categories, all recognisers exhibit similar high run time efficiency. The hybrid approach significantly outperforms the traditional systems as no — possibly invalid — assumptions about the distributions of feature vectors are made.

REFERENCES