DESCRIBING INTONATION WITH A PARAMETRIC MODEL

Gregor Möhler
moehler@ims.uni-stuttgart.de

Institute of Natural Language Processing, University of Stuttgart
Azenbergstr. 12, 70174 Stuttgart, Germany

ABSTRACT

In this study a data-based approach to intonation modeling is presented. The model incorporates knowledge from intonation theories like the expected types of F_0 movements and syllable anchoring. The knowledge is integrated into the model using an appropriate approximation function for F_0 parametrization. The F_0 parameters that result from the parametrization are predicted from a set of features using neural nets. The quality of the generated contours is assessed by means of numerical measures and perception tests. They show that the basic hypotheses about intonation description and modeling are in principle correct and that they have the potential to be successfully applied to speech synthesis. We argue for a clear interface with a linguistic description (using pitch-accent and boundary labels as input) and discourse structure (using pitch-range normalized F_0 parameters), even though current text-to-speech systems usually still do not have the capability to predict most of the appropriate information.

1. INTRODUCTION

Every model trained on data will improve when the right assumptions about the underlying phenomena can be made. For the modeling of intonation we are therefore examining intonation theories to use their main findings as a priori knowledge for our approach. Since the data used here has been analyzed according to the Tone-Sequence-Model we will start by summing up the main properties of this established theoretical background.

Originally introduced by Pierrehumbert the Tone-Sequence-Model (TSM) has been adapted to many other languages for the description of intonation [1]. The basic categories of the TSM are pitch accents and phrasal tones. All of them are based on the two atoms H and L, which are phonetically realized as high or low targets in the speaker's pitch range. The targets are related to the stress bearing units of the utterance (i.e. syllables in the case of germanic languages). In Pierrehumbert's notation the * denotes the accented syllable. In an L*+H accent, for example, a low target on the accented syllable is followed by a rise up towards a high target, which is usually reached within the next syllable.

In this study we are using the TSM adaptation of Féry, who modified the original tone inventory in order to capture the particular aspects of German intonation [2]. She found that German intonation can be described by 5 pitch accents and 2 boundary tones. The main pitch accents are an F_0 rise (L*+H), fall (H*+L) and rise-fall (L*+HL). Beside these common pitch accents she observed rare occurrences of an early peak and the so-called 'stylized contour' which can be found in vocatives. The intonation characteristics of at the boundaries can be described by a default intermediate boundary, a default phrase boundary (both have no explicit notation in Féry's work) and a high boundary tone (H%). The default boundaries essentially extend the last tone of the phrase-final pitch accent to the end of the phrase: A default boundary after L*+H is therefore interpreted as high and a default boundary after H*+L as low. The additional high boundary tone H% is necessary to explain a final rise after a falling accent.

While there are commonly used models for the description of pitch movements, no widely accepted model exists for the pitch range. Usually the pitch range of an utterance is defined by the upper and lower boundaries of the F_0 contour. Problems occur if an accent is subject to expressive raising which places the high target of the accent beyond the pitch range of the utterance [3]. Classifying expressive raising is very subjective, which makes it hard to determine the pitch range automatically. As no rules are known we will apply a statistical method to determine the pitch range of an utterance (see section 3.3).

While the F_0 movements represented by the TSM elements are in accordance with local semantic functions like focus and topic [3], the pitch range of the F_0 contour is related to the higher level structure of discourse. Even though it is very hard to derive semantic and higher order linguistic analysis from a written text it is important to take these influences into consideration by providing an appropriate interface with the intonation model. We can expect that semantic information will be at least partially available in the future. This is especially true for speech-based automatic translation and dialog systems where the modules prior to synthesis operate on the basis of semantic information. A clear interface with this level of linguistic description will facilitate the future incorporation of speech synthesis into these systems.

2. DATABASE DESCRIPTION

The corpus used for this study consists of 72 news stories recorded from the digital satellite radio (DSR). They were read by a male professional news reader. The total length of the corpus is 48 minutes.

Word, syllable and phoneme transcriptions of the corpus were added using methods of forced alignment [4]. The manual prosodic annotation is based on the tone sequence model for German described above. It has been modified in some points to meet the requirements of labelling (e.g. an explicit notation of the default boundaries has been added). The development of the used label system is part of the GtoBI (German Tone and Break Indices) project [5].

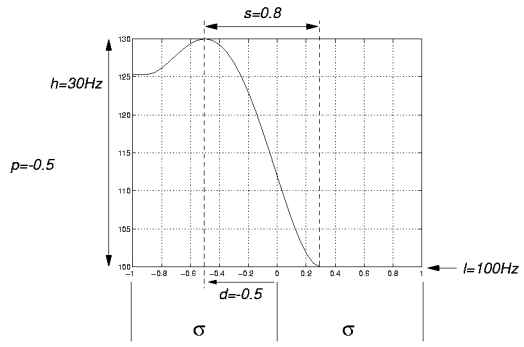

Figure 1: The parameters of the approximation function: *p* (movement type), *s* (steepness), *d* (alignment), *l* (base level), *h* (amplitude).

3. METHODS

The data-based intonation model presented here relies on a set of parameters which is automatically extracted from the F_0 contour. The parametrization is based on a specially designed function which is approximated to the F_0 curve. The free parameters of the function are found during the approximation. They describe the movement of the F_0 curve. In a synthesis application these parameters would be predicted and the F_0 curve is constructed on the basis of the underlying function.

3.1. Design of the approximation function

We use a prior knowledge of the expected F_0 movements to design the approximation function. It is based on the general findings of the Tone-Sequence Model. Hence, the function should be able to model F_0 movements that are either rising or falling or of the rising-falling type. Since syllables are the stress bearing units we want the approximation to be syllable based.

The approximation function is a 4th degree polynomial function that is extended by constant lines in its turning points (cf. Fig. 1). It is defined within the interval $-1 \leq p \leq 1$. During the parametrization process the accented syllable and the post-accent syllable are normalized and mapped onto the intervals [-1;0] and [0;1] respectively. This normalization reflects the syllable-based anchoring of the targets in the TSM. The approximation function can be shaped by a set of 5 parameters. They have been chosen to allow a meaningful interpretation of the underlying movement.

The function moves within the base level *l* and the maximum *l*+*h*. The parameter *l*, therefore, symbolizes the potential place of the L target, and *h* represents the amplitude of the movement.

The parameter *p* represents the basic shape of the movement. It follows the principles of the *tilt* parameter established in [6]: A value of p=1 represents a pure rise, p=-1 results in a pure fall and for p=0 a rising-falling contour is modeled. But the parameter *p* is not restricted to these discrete values. For $-1 \leq p \leq 1$ any movement that consists of a rising part followed by a falling part can be modeled (cf. Fig. 2). A falling-rising shape is not part of the movements accepted within the TSM and can therefore not be modeled with this function.

The parameter *d* indicates the alignment of the peak point within the two-syllable window. And, finally, the parameter *s* describes the time within which the movement rises or falls between the levels of *l* to *l*+*h*.

3.2. Approximation process

For the process of approximation the F_0 contours were median-smoothed, interpolated through unvoiced periods and segmented in two-syllable windows overlapping by one syllable. The time scale within this window was normalized according to the syllables' lengths. We applied the Nelder-Meade simplex-search method for the approximation. Different sets of starting values were applied to find the best approximation. When the algorithm did not converge we restricted the approximation on the first half of the window (accented syllable). This situation occurs e.g. when 2 accents follow each other very closely, resulting in F_0 movements that cannot be described by the approximation function.

We parametrized the complete news corpus. To assess the quality of the parametric description we reconstructed the F_0 contours based on the five parameters. The overlapping parts of the predicted movements were averaged, the resulting contours smoothed with a low pass filter of 33 Hz. As a numerical quality measure we calculated the root mean squared error (RMSE) and the correlation between the resulting contours and their original counterparts. We achieved a RMSE of 6.8Hz and a correlation of 0.94. These findings show a very close approximation of the F_0 contours by the model. Visual comparison of the F_0 contours and informal listening tests underlined the numerical results, so that we did not carry out any further (e.g. perceptual) evaluations.

3.3. Pitch range estimation

Usually, two lines represent the upper and lower boundary of the pitch range. We also intend to follow this approach. The upper boundary of the pitch range is defined by the highest peak in the utterance [3]. Pitch range estimation takes place after the param-

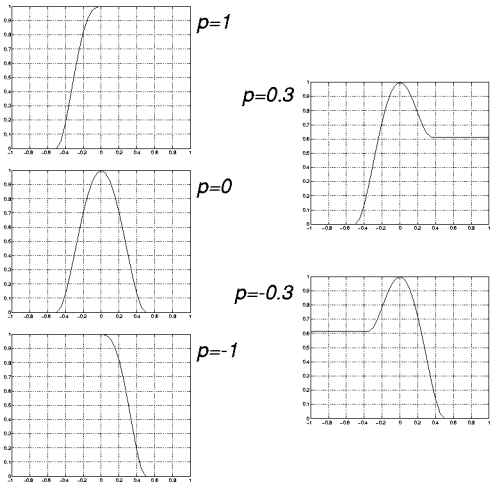

Figure 2: The parametrization function for different values of the shape parameter *p*, ranging from a pure rise to a pure fall

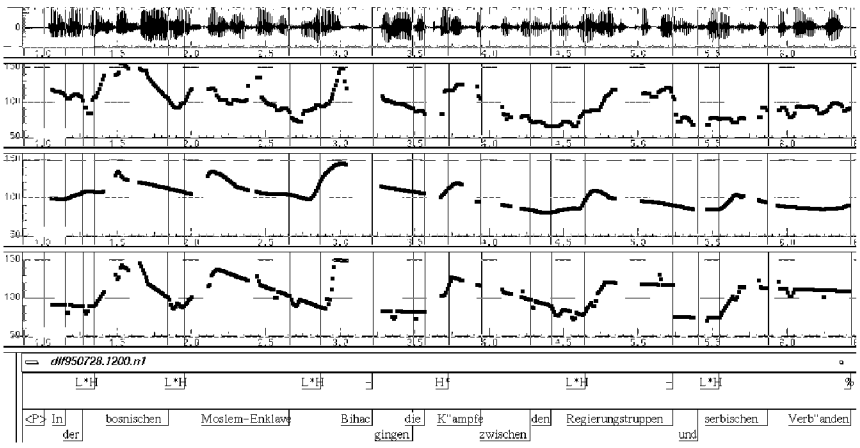

Figure 3: Example utterance: Speech signal, original F_0 contour, F_0 contour generated from data using the model presented in this paper, F_0 contour generated by rules, prosodic and word transcription.

etrization of the F_0 contours. For this reason we look at all F_0 parametrizations of the pitch accents within an intonation phrase. The maximum of all values, $l+h$, serves as the pitch range's upper boundary. The effect of expressive rising is taken into account by excluding outliers from this statistics. The same principle is applied to all l parameters, resulting in an estimation of the pitch range's lower boundary.

3.4. Prediction of F_0 contours

We used bidirectional recurrent neural nets (BRNN) to predict the parameters from a set of features [7]. Recurrent nets are superior to multi-layer perceptrons (MLP) when dealing with time correlated data. Time dependencies are captured by the state neuron layer which is fully connected to the input neurons and contains a feed-back loop.

The parametric F_0 description of each syllable in the utterance makes up the output vectors of the neural net. To establish an interface with a separate pitch range model we normalized the parameters l (base level) and h (amplitude) with the pitch range values of the respective utterance. Therefore the F_0 contours will be predicted in a normalized pitch range and are only later translated into an actual pitch contour using a specific pitch range value.

For every syllable a set of features was extracted: type of the accent and boundary (GToBI label) determined for a window of 5 syllables around the syllable in question. Other features include the distance to the preceding and following accent, the length of the intonation phrase, the position within the intonation phrase and the length of 3 specific elements defined within the syllable: the stable, pre-stable and post-stable part as motivated by the work on pitch perception by House [8].

We trained the BRNN on 60 of the 72 news stories. Both input and output data had been normalized to a mean of zero and a variance of one. The best topology could be found for 6 state neurons in forward direction and 4 neurons in backward direction.

From the predicted parameters we reconstructed the F_0 contours of the 12 news stories that make up the test corpus (see Fig. 3). The F_0 curve within syllables without pitch accent or boundary was not reconstructed from the predicted F_0 parameters. Instead we only used a single point (the mean of the predicted movement) in these syllables. We did so, because only syllables with pitch accents or boundaries exhibit F_0 movements that can be consistently modeled by our function. Unmarked syllables, however, are characterized by global movements which are basically an interpolation between the pitch accents .

The numerical evaluation of the predicted F_0 contours in the test set resulted in an RMSE of 16.0 Hz and a correlation of 0.64.

4. PERCEPTUAL EVALUATION

From the 12 news stories in the test set we extracted 7 intonation phrases for the perception experiment. The stimuli were generated using speech synthesis with the publicly available diphones from the MBROLA project [9]. We synthesized the phonetic transcription directly from the database. The phone duration of the stimuli is therefore an immediate result of the aligner's output.

Three versions of each intonation phrase were generated: One with the original F_0 contour, one with the F_0 contour generated by the model described in this paper and one with an F_0 contour generated by means of a rule-based intonation generation [10]. The rule-based generation is based on a direct mapping of TSM targets into F_0 targets according to rules operating on the syllable structure. The resulting F_0 contours are appropriate phonetic representations of the underlying TSM description, although not always completely natural due to the restricted number of rules involved.

The stimuli were presented in a web-based test environment, where the raters could click on icons to listen to the stimuli. To rate the stimuli, they pressed the respective answer buttons resulting in the automatic presentation of the next stimulus (or pair of stimuli). Thus we had no direct control over the timing of the test. All stimuli were repeated once.

Experiment 1. In the first perception test we asked for an evaluation of the naturalness of the intonation of the stimulus presented. We wanted to know whether our listeners would accept the intonation contours as possible realizations of some underlying meaning. We also used stimuli with completely flat intonation as control items. The listeners could answer on a 6 step scale ranging from 0 to 5. All ratings had verbal attributes (like very natural, almost natural, etc.) assigned within brackets.

The original contours were rated with an average score of 3.39, rule-based contours with 3.09 and the data-based generated contours with 3.79. Original and rule-based contours showed no significantly different ratings. whereas the data-based approach was rated significantly higher (variance analysis on all versions, $p< 10^{-5}$ and Turkey-Test, $p<10^{-3}$) than all other methods. The test item with flat intonation was rated with 1.95, which is significantly lower than all other stimuli. We can see that the neural net based contours are rated to be very natural. They even got better results than the originals. However, we don't wish to assign too much importance to the lower rating of the original contours, because it might be due to their unsmoothed contours as compared to the smoother generated contours.

Experiments 2 and 3. In two experiments we compared stimuli with original intonation contours and stimuli with rule-based (experiment 2) or data-based (experiment 3) generated contours. We asked the listeners to evaluate the difference between the two intonations on a scale from 0 to 4 (with verbal attributes) . This test gives us an interpretable result under the assumption that the original contours are a prototypical realization of the labelled TSM notation. If the two intonation contours are rated to be very similar, we can conclude that the two stimuli have the same interpretation (as far as it is expressed by intonation). As control pairs we also provided two stimuli with accents that were distinctly different from each other (different pitch accent types).

The rating for the rule-based version was 1.28 and 1.81 for the data-based version. These two results are significantly different (variance analysis, significance$<10^{-3}$). We analyzed the bad results of the data-based version and found that the two examples rated with a bad score had problems with modeling a high boundary tone and with the alignment of a falling accent. High boundaries are rare in the news corpus, which might explain the difficulties. However, there seem to remain problems due to the inexact alignment of accents. The deviation from the original intonation contours was nevertheless significantly smaller than for the test stimuli.

Experiment 4. We wanted to know whether listeners would prefer rule-based stimuli or stimuli with intonation generated by neural nets. Thus we presented these stimuli in pairs and asked which one listeners preferred. It turned out that for all but one stimuli presented listeners chose the data-based model. We can conclude that the model presented in this study is preferable to a direct rule-based implementation of intonation modeling.

5. DISCUSSION

We presented a data-based method of intonation generation. It successfully uses knowledge known taken from intonation theory in the parametrization step. The F_0 contours generated by the neural nets were rated as highly natural. The perception test also revealed that there is some potential for improvement in the alignment of accents. A timing model as presented by [11] might be helpful here. In general, however, we can say, that our hypotheses about intonation description, intonation modeling and fundamental frequency generation are in principle correct. The intonation model has been fully integrated into the FESTIVAL text-to-speech system [12].

As the F_0 parametrization has been developed to code knowledge about intonational phenomena, it can also be used as a tool for phonetic analysis. It has been successfully applied to the analysis of the Lithuanian accent [13] and to the register analysis of discourse prosody [10].

REFERENCES

[1] Pierrehumbert, J. (1980). *The Phonology and Phonetics of English Intonation*. Ph.D. Thesis. MIT, Cambridge, MA.

[2] Féry, C. (1993). *German Intonational Patterns*. Niemayer, Tübingen.

[3] Mayer, J. (1997). *Intonation und Bedeutung, Aspekte der Prosodie-Semantik-Schnittstelle im Deutschen*. Dissertation, Universität Stuttgart.

[4] Rapp, S. (1995). Automatic phonemic transcription and linguistic annotation from known text with Hidden Markov Models / An Aligner for German. Proceedings of *ELSNET goes east and IMACS Workshop ``Integration of Language and Speech in Academia and Industry''*, Moscow.

[5] Benzmüller, R. & M. Grice (1997). *Trainingsmaterialien zur Etikettierung deutscher Intonation mit GToBI*. In: Phonus 3, Saarbrücken.

[6] Taylor, P. (1994). Synthesizing Conversational Intonation from Linguistically Rich Input. *Proc. of ESCA/IEEE Workshop on Speech Synthesis*, Mohonk, NY, 175-178.

[7] Schuster, M. & K. K. Paliwal (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing* 45 : 2673-2681.

[8] House, D. (1996). Differential perception of the tonal contours through the syllable. In: *Proceedings of ICSLP*, Philadelphia, 2048--2051.

[9] Dutoit, T. et. al (1996). The *MBROLA* project: Towards a set of high-quality speech synthesizers free of use for non-commercial purposes. In: *Proceedings of ICSLP,* Philadelphia, 1393-1396.

[10] Möhler, G (1998). *Eine theoriebasierte Modellierung der deutschen Intonation für die Sprachsynthese*. Dissertation, University of Stuttgart.

[11] van Santen, J. & B. Möbius (1997). Modeling pitch accent curves. In: *Proceedings of the ESCA Workshop on Intonation: Theory, Models and Applications*, Athens, 321-324.

[12] Black A.W. & P. Taylor (1997). *The Festival Speech synthesis System*. System Dokumentation, University of Edinburgh.

[13] Dogil, G. & G. Möhler (1998). Phonetic invariance and phonological stability: Lithuanian pitch accents. In: *Proceedings of ICSLP*, Sydney, this volume.