View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Open Access LMU

Text-based prediction of automatically extracted intonation contour classes

Uwe D. Reichel

Institute of Phonetics and Speech Processing
University of Munich, Schellingstr. 3, 80799 Munich,Germany
reichelu@phonetik.uni-muenchen.de

Abstract. In this paper classifiers for text-based prediction of intonation contour classes are compared. The contour classes were derived automatically by a method presented in Reichel (2006), and the following classifiers were utilised for prediction: Bayes classifier, C4.5 decision trees, perceptrons, and linear feedforward networks. Prediction accuracies amounted from 38.0% (Perceptron) to 66.6% (Linear Network).

1 Introduction

Given approaches of description of intonation can roughly be divided into symbolic, parametric and perception-based methods. They all have certain shortcomings: symbolic approaches like the tone sequence approach (Pierrehumbert, 1980) or the Kiel intonation model (Kohler, 1991) as well as parametric approaches like the Fujisaki model (Fujisaki, 1987) and PaintE (Moehler et al., 1998) generally require time consuming prosodic labelling and run the risk of low inter-labeller agreement and intra-labeller consistency, which reduces the amount of reliably annotated data (Grice et al., 1996). Furthermore, being applied to new languages the label inventory often needs adjustments (Reyelt et al., 1996). Also perception-based models like IPO (t'Hart et al., 1990) require perceptual readjustment for each new language (Adriaens, 1991). In Reichel (2006) we presented an alternative purely data-driven approach to extract intonation contour classes. We distinguished local classes spanning over a short syllable window and global classes intended to correspond to intonation phrases. Global and local contour classes are superposed like in Fujisaki (1987) to form the final f0 contour. In perception tests carried out in the same study comparing stimuli with original and resynthesised f0 contours, the resynthesised contours were judged as less natural but functionally equivalent to the original contours.

In this paper we address the question whether the classes could be utilised in text-to-speech synthesis (TTS). Since TTS generally requires text-based predictability of the f0 contour, here we describe first attempts to predict the contour classes from text, concentrating on global classes. For this purpose we utilised some standard classifiers as decision trees and neural networks and compared their performances.

2 Data

For contour class extraction and text-based prediction we used parts of the IMS radio news corpus (Rapp, 1998) containing news text read by a professional male speaker. The data we used comprises 3985 syllables (about 14 minutes) and is amongst others segmented on the phone and syllable level. F0 values were extracted with a sample frequency of 100 Hz using autocorrelation implemented in *Praat* and transformed to semitones. Voiceless segments were bridged by cubic spline interpolation and the contours were smoothed using the Savitzky-Golay filter (order 3, window length 5). Data size is given in table 1.

Table 1. Data size.

	Tokens	Types
global contour segments	708	6
POS labels	1869	37

3 Extraction of intonation contour classes

In the next paragraphs the method for contour class extraction is introduced in short. A more detailed description can be found in Reichel (2006).

3.1 Local contour classes

Local classes are derived in an iterative manner starting at the syllable level. The smoothed and time normalised f0 contours are stylised by polynomes, and classes are derived by Kmeans clustering of the polynomial shape coefficients (ignoring the vertical offset). Neighbouring contour segments are merged if the associated classes are interdependent, and the stylisation and clustering process restarts with the new found larger segments. The process terminates if no further segment merging is required by any systematic class co-occurrence. Local contour classes derived by this method are presented in Figure 1.

3.2 Global contour classes

To extract global contour classes the utterances are split up into segments very roughly corresponding to intonation phrases (IP). The segmentation is guided by speech pauses and pitch discontinuities. Global contour classes are then derived by linear stylisation of the f0 baselines given in the segments and clustering of the resulting shape coefficients (again ignoring the vertical offset). The derived global classes are shown in Figure 2.

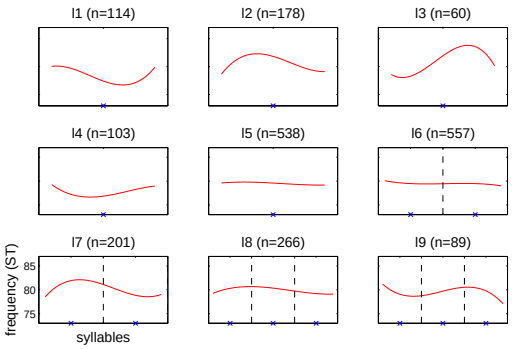

Fig. 1. *Local contour classes. All contours are shifted to the mean of 80 ST. Time is normalised. Syllable boundaries are marked by vertical dashed lines, nucleus centres by crosses.*

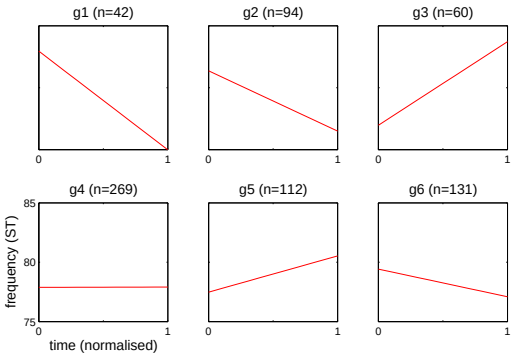

Fig. 2. *Global contour classes (declination baselines). Time is normalised to the interval* [0 1].

3.3 Resynthesis

The final f0 contour is calculated by superposition of global and local contours (see Figure 3). For each syllable an f0 register is derived from the global contour class, the f0 starting point within the IP and the syllable position within the IP. The local contour classes are adjusted to the durations and the constituent structures of the concerned syllables and added on the registers.

The f0 starting points for the IPs are derived from predicted pitch resets. For pitch reset prediction the following multiple linear regression model is utilised:

$$r = -5.08 \cdot s_1 - 5.17 \cdot s_2 + 5.14 \cdot d - 5.61 \cdot b,$$

where s_1 and s_2 are the slope coefficients of the preceding and the following global contour, d is the pause duration between the IPs, and b is the register of

the boundary preceding syllable. The root mean squared error between original and predicted resets amounts 2.55 semitones and Pearson's r is 0.68.

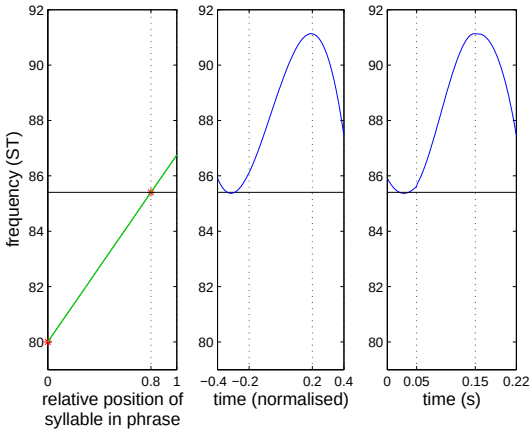

Fig. 3. *Combination of global and local contour.* **Left:** *The segment's register baseline is predicted by original frequency offset (here: 80 ST), global contour associated to corresponding global contour class (here: class g_3), and relative position of the segment within the intonational phrase (here: 0.8).* **Middle:** *The baseline value of the local contour given here by local contour class l_3 is shifted to this value.* **Right:** *The contour is aligned to the original time ranges of syllable onset (0s–0.05s), nucleus (0.05s–0.15s) and coda (0.15s–0.22s).*

4 Text-based prediction of global contour classes

In this study the prediction is restricted to global contour classes, and the segmentation is assumed to be given. The prediction of the segment boundaries needed for TTS applications could be carried out by relating them to part of speech (POS) sequences (e.g. Black et al., 1997).

4.1 Features

Prediction is based on the following feature pool:

- intonation class dependent probabilities of POS sequences
- phrase lengths in syllables and words

POS probabilities For each contour class C probabilities for part of speech sequences S are estimated by linear interpolated uni-, bi-, and trigram language

models using Good-Turing smoothing. The high relative entropy values between the resulting class dependent probability distributions could indicate good pairwise separability but also data sparseness. They are shown in table 2.

Table 2. Relative Entropies of probability distributions P_x and P_y, x and y representing intonation classes. Each cell shows the mean extra number of bits needed to encode the POS sequence S_x following the distribution P_x (rows) with the distribution P_y (columns).

	P_{g1}	P_{g2}	P_{g3}	P_{g4}	P_{g5}	P_{g6}
$P_{g1}(S_1)$	0	548.87	663.16	475.94	599.10	472.32
$P_{g2}(S_2)$	1248.35	0	1371.20	940.51	1189.46	919.92
$P_{g3}(S_3)$	580.07	514.30	0	386.28	425.29	387.44
$P_{g4}(S_4)$	2488.15	1840.35	2685.81	0	2253.32	1705.90
$P_{g5}(S_5)$	1054.47	850.03	914.49	661.48	0	789.10
$P_{g6}(S_6)$	1613.31	1299.49	1788.61	1128.50	1681.52	0

Other features Table 3 shows the mean phrase lengths in syllables and tokens for each of the intonation classes.

Table 3. Mean phrase lengths for each intonation class c.

Contour class	n(syllables)	n(POS tokens)
c1	7.71	5.14
c2	7.64	4.88
c3	4.45	3.45
c4	3.54	3.41
c5	4.38	3.55
c6	6.45	4.81

4.2 Classifiers

We tested the following classifiers for the prediction task:

- Bayes
- C4.5 decision trees
- Perceptrons
- Linear feedforward networks

Bayes Within the framework of Bayes classification the predicted contour class $\hat{C}$ is the class which maximises the following expression:

$$\hat{C} = \arg\max_{C} \left[P(C|S) \right]$$
$$\hat{C} = \arg\max_{C} \left[P(S|C) \cdot P(C) \right],$$

where S is the POS sequence given in the examined segment, and $P(S|C)$ is estimated as described in section 4.1 by linear interpolated n-gram models.

Extensions In equivalence to approaches like in Timoshenko et al. (2006) the maximum probability decision of Bayes can be replaced by a decision of another classifier fed by conditional probabilities $P(C|S)$ for all classes C given a POS sequence S. This classifier can further be provided with additional knowledge.

Other classifiers The other classifiers tested in this study are C4.5 decision trees, perceptrons and feedforward networks. Two feature pools were used, one containing the $P(C|S)$ for all six contour classes C, the other additionally containing phrase length given by the number of words.

For the one-layer perceptron the hard limit transfer function and the perceptron weight/bias learning rule were chosen

Two classes of feedforward networks were trained, one with and the other without a hidden layer. The input layer contained one unit for each feature, the output layer one unit for each class. The hidden layer comprised the same number of units as the input layer. Saturating linear transfer functions were uniformly used. The gradient descent weight/bias learning function was chosen in Levenberg-Marquardt backpropagation training. The networks output was binarised by setting the highest activity to one and the others to zero.

Baseline model All trained models were tested against each other and against a baseline model consisting of chance guesses guided by class occurrence probabilities.

Procedure To evaluate each classifier we applied 30-fold cross-validation each time using 80% of the data for training and 20% for testing. The subsequent tables refer to the derived mean performances.

4.3 Results

As can be seen in table 4 the feedforward networks yielded the best results. The winner was the feedforward network without hidden layers using all $P(C|S)$s and phrase length as features. The perceptron yielded the worst performances. All classifiers outperformed the baseline model.

Table 5 shows which performance differences are significant. Feedforward nets without hidden layers significantly outperform all other classifiers.

Table 4. Performances (in %) in dependence of classifier and feature pool

	P_C	+n(tokens)
Bayes	56.57	–
Perceptron	38.02	37.53
Feedforward, 2 layers	64.71	66.60
Feedforward, 3 layers	58.97	63.26
C4.5	58.77	58.68
BL	22.89	

Table 5. Significance of performance differences. $*(*)$: row classifier performs (highly) significantly better than column classifier (one-tailed t-test for matched samples). 'FF': feedforward, H: hidden layer, 'l': using phrase length feature, DT: decision tree, B: Bayes, P: perceptron, BL: baseline

	performs better than:									
	FFl	FF	FFHl	FFH	DT	DTl	B	P	Pl	BL
FFl			**	**	**	**	**	**	**	**
FF				*	**	**	**	**	**	**
FFHl					**	**	**	**	**	**
FFH							*	**	**	**
DT							*	**	**	**
DTl							*	**	**	**
B								**	**	**
P										**
Pl										**
BL										

5 Discussion

5.1 Interpretation of poor performances

Generally, the accuracies yielded in this study, the highest still below 70% are rather low. Possible reasons for this finding are:

- The data-driven derived contour classes could not be interpreted linguistically.
- The choice of features to predict the classes may be inadequate to some extent.
- The data may be too sparse for successful signal-based intonation class extraction and/or text-based class prediction.

At present the first hypothesis cannot be verified, since the other two explanations may also hold.

In this study very low-level features as POS labels and phrase lengths were used. Since intonation is affected by higher linguistic levels as syntax (e.g. Abney, 1991) and semantics (e.g. Mayer, 1997), the predictions may be based on insufficient linguistic analysis.

The data sparseness hypothesis is supported by the findings that feedforward networks with a hidden layer perform highly significantly worse than more simple feedforward networks, which indicates the tendency of overadaption appearing when the model size gets to large compared to the amount of training data. Also the high relative entropies between the class dependent POS probability distributions (see table 2) could indicate too small sample sizes.

Therefore it cannot be decided at the present state whether the contour classes are sufficiently related to linguistic concepts or not.

5.2 Extending Bayes Classification

Except for the perceptrons all other classifiers outperform Bayesian classification. This finding supports the approach not just to decide for the highest class probability but to leave the class decision to a classifier which has the knowledge about all class probabilities.

6 Conclusion and future directions

In this study a first attempt was made to predict automatically derived global intonation contour classes from text. More training data and a more elaborated linguistic analysis would probably further increase prediction performance. The next step will be to extend text based prediction for local contour classes.

References

1. Abney, S. (1991). Parsing by chunks. In: Berwick, R., Abney, S., and Tenny, C. (eds.), *Principle-Based Parsing*. Kluwer Academic Publishers, Dordrecht.
2. Adriaens, L.M.H. (1991). *Ein Modell deutscher Intonation: eine experimentell-phonetische Untersuchung nach den perzeptiv relevanten Grundfrequenznderungen in vorgelesenem Text.* Ph.D. thesis, University of Technology, Eindhoven.
3. Black, A.W. and Taylor, P. (1997). Assigning phrase breaks from part-of-speech sequences. *Proc. Eurospeech*, Rhodes, pages 995 – 998.
4. Fujisaki, H. (1987). A note on physiological and physical basis for the phrase and the accent components in the voice fundamental frequency contour. In O. Fujimura, editor, *Vocal physiology: voice production, mechanisms, and functions*, pages 165–175. Raven, New York.
5. Grice, M., Reyelt, M., Benzmueller, R., Mayer, J., and Batliner, A. (1996). Consistency in Transcription and Labelling of German Intonation with GToBI. In *Proc. ICSLP*, pages 1716–1719, New Castle, Delaware.
6. Kohler, K. (1991). A model of German intonation. In *AIPUK*, volume 25, pages 295–360. Kiel.

7. Mayer, J. (1997). *Intonation und Bedeutung: Aspekte der Prosodie-Semantik-Schnittstelle im Deutschen*. PhD thesis, IMS, Stuttgart.
8. Moehler, G. and Conkie, A. (1998). Parametric modeling of intonation using vector quantization. In *Proc. 3rd ESCA Workshop on Speech Synthesis*, Jenolan Caves, Australia.
9. Pierrehumbert, J. (1980). *The phonology and phonetics of English intonation*. Ph.D. thesis, MIT, Cambridge, MA.
10. Rapp, S. (1998). *Automatisierte Erstellung von Korpora fuer die Prosodieforschung*. Ph.D. thesis, IMS, Universitt Stuttgart.
11. Reichel, U.D. (2006). Data-driven Extraction of Intonation Contour Classes. In *Proc. 6th ISCA Workshop on Speech Synthesis*, Bonn, Germany.
12. Reyelt, M., Grice, M., Benzmueller, R., Mayer, J., and Batliner, A. (1996). Prosodische Etikettierung des Deutschen mit ToBI. In Gibbon, D., editor, *Natural Language and Speech Technology, Results of the third KONVENS conference*, pages 144–155. Mouton de Gruyter, Berlin, New York.
13. t'Hart, J. and Collier, R. and Cohen, A. (1990). *A perceptual study of intonation: An experimental-phonetic approach to speech melody*. Cambridge University Press, Cambridge.
14. Timoshenko, E. and Bauer, J.G. (2006). Language Identification using Unsupervised Model Training. In *Proc. AST*, pages 159–164. Maribor.