Algorithmic classification of pitch movements

Louis ten Bosch
Institute for Perception Research/IPO
P.O. Box 513, 5600 MB Eindhoven, the Netherlands.
e-mail: tenbosch@prl.philips.nl

ABSTRACT
In this paper, we discuss the construction of an algorithm that classifies pitch movements according to the IPO intonation labelling. The classification is performed by a feed-forward network, interpreted as a multi-linear classifier. In speaker-independent tests on a corpus of speech read by non-professionals, up to 81 % of the 279 pitch movements in the test corpus were correctly classified. These results are obtained by using information from the sampled speech data files only; a grammar will be used in the second stage of this study.

Keywords: Automatic classification, multi-linear classification, IPO-intonation system, speech recognition.

1. INTRODUCTION.

In this paper, an algorithm will be described aiming at the (semi-)automatic classification of pitch movements. The algorithm is trained and tested for Dutch. Its input is a sampled data file of an utterance; its output consists of a character string containing intonation transcriptions ('labelling'). Optimally, the algorithm should come up with a labelling that is indistinguishable from transcriptions produced by human intonation experts.

Algorithms that classify elementary patterns of speech melody are useful, e.g., the detection of phrase boundaries and accented syllables, the filtering of acoustically based hypotheses from an ASR-algorithm (Ostendorf, Wightman and Veilleux, 1991; Wightman and Ostendorf, 1992), and the labelling of large speech corpora.

In the present approach, the intonation labelling convention will be used which is known as the IPO-labelling. This labelling is chosen due to the relations posed in the theory between acoustic realization and perceptual labels ('t Hart, Collier, and Cohen (1990). This system defines ten labels (five different pitch rises labelled '1' to '5', five falls labelled 'A' to 'E'), with additional labels referring to a 'pointed hat' ('P'). Syllables bearing a perceptually relevant pitch movement can be labelled with at most one of these labels. Five labels are most common: '1', '2', 'A', 'B', and 'P' (also denoted '1&A'). The functional difference between the rises '1' and '2' and falls 'A' and 'B' corresponds to a phonetic difference with respect to the exact timing of the pitch movement: Accent-lending movements such as '1' and 'A' are generally earlier in the syllable than are the non-accent-lending movements '2' and 'B'.

The 'IPO-intonation-grammar' prescribes the permitted sequences of labels within one utterance ('t Hart *et al.*, 1990).

2. DESIGN OF THE ALGORITHM.

Several attempts have been made to come to a (semi)automatic classification of pitch movements, using, e.g., dynamic programming (Brew and Isard, 1990), or Hidden Markov Modelling (Butzberger, 1990). These approaches have proven to be fairly successful. We did not opt for these approaches, however, due to the difficulty of a proper interpretation of the many model parameters.

In this study, classification is based on multi-linear discrimination on features extracted from the sampled data file. The algorithm consists of two steps: (a) training, which is based on a labelled training corpus, and (b) classification, based on multi-linear discrimination.

(a) Training.

For the construction of a labelled database, a number of 800 Dutch sentences have been recorded. The average number of words per sentence was 7.4. These sentences (elicited speech) were spoken by over forty different speakers, male as well as female.

Sentences were manually labelled according to the IPO-system by four expert intonologists independently. A common subset was labelled, on which a consensus labelling has been defined. Eventually, a resulting total of 249 sentences were used in the test described below. The total number of labelled syllables was 817 (an average of 3.3 labelled syllables per sentence).

The distribution of the labels over the set of labelled syllables is '1' 31 %, '2' 14 %, 'A' 15 %, 'B' 14 %, 'P' 16 %, and other 10 %.

For each label, a corresponding class of acoustic realizations was constructed. The data space was constructed by feature representation in four steps (cf. Ten Bosch, 1993): (1) Pitch determination, (2) Correction of pitch measurements, (3) Determination of vowel onsets, and (4) Choice of pitch reference points.

Step (2) is included since the pitch determination algorithm usually returns the correct pitch as perceived on a (sub)syllabic scale. The actual pitch as it is perceived on the sentence scale (without gating) may deviate from the PDA outcome. A reinterpretation of the pitch contour results (figure 1).

In step (4), five pitch measurements per syllable were chosen as reference measurements: two measurements in the previous syllable, two measurements in the current syllable, and one measurement in the next syllable. These measurements were anchored at the moments of vowel onsets (Ten Bosch, 1993). The resulting data set (denoted $\mathcal{D}$) has dimension 5.

We make two observations about this representation. It is rather 'poor' in the sense that it does not make use of other spectral features. However, this 'poor' representation is sufficiently rich to cover the main distinctive features between the label classes (see below). Secondly, it deviates from the more standard representation spanned by 'excursion size', or moment of start and end of a pitch movement. The results of the training step allow these 'classical' features to be used in a description of class prototypes, but these features are certainly not unique.

(b) Classification.

The classification training was done on a subset of 65 % of the available set of labelled syllables. Most label classes in $\mathcal{D}$ are convex; they however do not necessarily obey a gaussian distribution (Ten Bosch, 1993). Consequently, the design of a Bayes classifier is not straightforward, and the recognition technique that is based on nearest prototypes may require more than one prototype per class (Ullmann, 1973, chapter 4). For the classification, a multi-linear discrimination was applied (cf. Fukunaga, 1972). The actual implementation of this optimization is done by a multi-layer classifier, i.e. a multi-layer perceptron (MLP), provided with a 5-n_h-n_y-topology, n_h ($2 \leq n_h \leq 5$) and n_y denoting the number of hidden units, and the number of output categories, respectively. For small-sized topologies, the MLP-results can be interpreted in a precise manner by relating them to a posteriori probabilities and CART-node questions (Richard and Lippmann, 1991; Breiman *et al.*, 1984).

In table 1, a summary is given of the results. The table shows results for several values of n_h and n_y. The normalized error (norm. error) denotes the mean error at an output unit. The column 'class. rate' denotes the fraction of correctly classified pitch movements. To cross validate the minimization, it was performed on 65 % of the available data, and tested on the remaining 35 % (279 syllables).

The results can be interpreted as follows. If the number of output categories n_y is clamped to 2, the best two 'class groupings' are {A, B} and {1, 2, P} (first row in table 1). Here, 'P' is more likely to belong to the group {'1', '2'}, rather than to {'A', 'B'}. An increase in the number of hidden units n_h, i.e. of the number of separating hyperplanes L_i used in the multi-linear discrimination in $\mathcal{D}$, shows an increasing classification rate (class assignment). If each class is to be labelled separately, an acceptable value of n_h is 5, as can be seen from the last four rows. As is suggested by the last row, it does not make sense to increase the discriminative power within $\mathcal{D}$ in order to optimize the classification rate substantially.

The result presented in the last row is tentative. It is the best result among 35 minimizations with substantially different initializations for the positions of L_i.

3. DISCUSSION.

In this paper, an attempt has been made to classify pitch movements by multi-linear discrimination. The main results are presented in table 1. These results have a unique interpretation from the point of view of technical optimization. The data obtained so far suggest that *unique* prototypical acoustical realizations of a class do not exist. In other words, prototypes do not specify the class topology on their own. Only under certain conditions, such as equal covariance matrices for each of the classes, the linear classification can be translated into a prototypical approach. These conditions are likely not to hold in $\mathcal{D}$. The present approach allows to look for distinctive features in the form of a set of hyperplanes L_i in $\mathcal{D}$, each hyperplane representing a specific 'property', i.e., a linear combination of the input features.

The question of how the classification results can be lined up with the 'classical' results in 't Hart *et al.* (1990) is solved by a close examination of the resulting MLP-weights. This shows that the difference between, e.g., '1' and '2' is mainly due to the value of the syllable-initial pitch in the current syllable relative to the syllable-final pitch in the previous and the current syllable. The behaviour of the 'classical' parameters that were known to be class-specific (e.g. timing and excursion differences, see 't Hart *et al.*, 1990) could be traced back in the test data as a trend only. This suggests that 'higher order' prosodic information (accents, grammar) must be used to further disambiguate between '1' and '2' or between 'A' and 'B'.

A final remark deals with the use of an intonation grammar. The disambiguation capability of the grammar presented in 't Hart *et al.* (1990) is, on the basis of the database presently used, for {1, 2} and {A, B} estimated to be 0.3. This means that the grammar disambiguates in the questions of the tree nodes {1, 2} and {A, B} in about one third of these cases.

ACKNOWLEDGEMENTS:

This research is supported by the Dutch Foundation for Scientific Research NWO, project no. 030-173-019.

REFERENCES:

Bosch, L.F.M. ten (1993). 'On the automatic classification of pitch movements.' Proceedings of the Eurospeech conference, Berlin, Germany.

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984). *Classification and regression trees.* Wadsworth, Belmont, CA.

Brew, C., and Isard, S. (1990). 'Principles of contour labelling.' DYANA: Dynamic interpretation of natural language. CSTR-report, Edinburgh, UK.

Butzberger, J.W. (1990). *Statistical methods for analysis and recognition of intonation patterns in speech.* Thesis, Boston university.

Fukunaga, K. (1972). *Introduction to statistical pattern recognition.* Academic Press, NY.

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

Ostendorf, M., Wightman, C.W., and Veilleux, N.M. (1991). 'Parse scoring with prosodic information: An analysis-by-synthesis approach.' Manuscript, Boston University.

Richard, M.D., and Lippmann, R.P. (1991). 'Neural network classifiers estimate Bayesian *a posteriori* probabilities.' Neural Computation 3, p. 461 - 483.

Ullmann, J.R. (1973). *Pattern recognition techniques.* Butterworth, London, UK.

Wightman, C.W., and Ostendorf, M. (1992). 'Automatic recognition of intonational features.' Proceedings 1992 International Conference on Acoustics, Speech, and Signal Processing, San Francisco. 221–224.

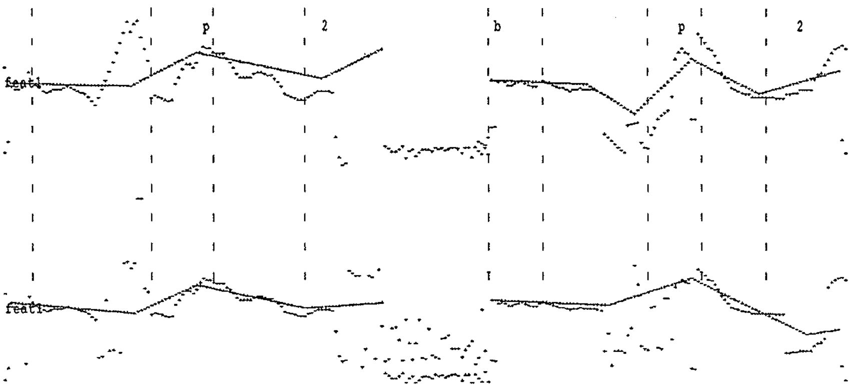

Figure 1. An example of correction and stylization of measured pitch. Along the abscissa, time is plotted. The pitch (scaled) is indicated along the ordinate. Correction and stylization remove perceptually irrelevant pitch jumps (e.g., octave errors).
Bottom: the original pitch measurement and its stylization.
Top: corrected pitch measurement and its stylization.

Table 1: *Results of the classification as a function of the two parameters n_h and n_y. For an explanation see the text.*

n_h	n_y	norm. error	class. rate	classes
2	2	0.81	0.79	{1, 2} ∪ {P, A, B}
2	2	0.54	0.85	{1, 2, P} ∪ {A, B}
3	2	0.59	0.83	{1, 2} ∪ {P, A, B}
3	2	0.33	0.89	{1, 2, P} ∪ {A, B}
3	3	0.41	0.81	{1, 2} ∪ {P} ∪ {A, B}
4	3	0.22	0.92	{1, 2} ∪ {P} ∪ {A, B}
4	5	0.42	0.54	{1} ∪ {2} ∪ {P} ∪ {A} ∪ {B}
5	5	0.27	0.81	{1} ∪ {2} ∪ {P} ∪ {A} ∪ {B}
6	5	0.23	0.83	{1} ∪ {2} ∪ {P} ∪ {A} ∪ {B}
≥ 7	5	< 0.23	$0.83 <$ rate < 0.86 ?	{1} ∪ {2} ∪ {P} ∪ {A} ∪ {B}