# AUTOMATIC CLASSIFICATION OF PITCH MOVEMENTS VIA MLP-BASED ESTIMATION OF CLASS PROBABILITIES.

Louis F.M. ten Bosch
Institute for Perception Research, P.O. Box 513, 5600 MB Eindhoven, The Netherlands

## ABSTRACT

In this paper, we study to what extent pitch movements in utterances can be classified automatically by using acoustical information and an intonation grammar. It will be shown that pitch movements can be classified into six categories with an agreement of about 80 percent compared with human transcriptions, on the basis of the pitch contour and the moments of vowel onsets. These six categories cover about 90 percent of all pitch movements used in the database (elicited speech). Results involving an intonation grammar are also presented.

## 1. INTRODUCTION

The automatic classification of prosodic events and the automatic transcription of pitch contours serves several purposes. First of all, in the field of automatic speech recognition (ASR), the automatic detection of accented syllables may be useful for reducing the size of hypothetical word lists during the generation of the acoustical hypotheses (cf. Bagshaw, 1993; Hieronymus, 1989). A more useful relation can be seen between the detection of boundary tones and grammatical constraints. A third application of prosodic information is the desambiguation of ambiguous utterances, especially in a (human–machine) dialogue setting.

In ten Bosch (1994), the relation has been studied between the statistical properties of acoustic realizations on the one hand, and the corresponding intonation category on the other. The intonation categories as used in this study were defined by 't Hart, Collier & Cohen (1990) (from here on referred to as HCC). In HCC, a general framework is presented for the description of Dutch intonation, in which the pitch contour of a well-formed Dutch utterance can be interpreted as a concatenation of different pitch movements. To this end, a collection of 'elementary' pitch movements is described, each of them associated with a different perceptual/phonetic 'category'. In essence, pitch movements are associated with (individual) syllables. Secondly, an intonation grammar was defined which describes the admissible sequences of categories for well-formed Dutch utterances.

In this paper, we will focus on the question how the pitch contour of an utterance can automatically be transcribed to be as close as possible to a transcription by intonation experts. This is done in two steps: (A) an algorithm is trained to separate and classify the most frequently occurring pitch movements occurring at individual syllables; (B) the pitch contour of an entire utterance is transcribed by applying a Viterbi algorithm on a probability lattice spanned by the syllables and the intonation categories. In the next sections, we first deal with preliminaries. After that, a few results concerning step (A) are briefly discussed; this material has been described in ten Bosch (forthcoming). Next, step (B) is focussed on in more detail.

## 2. PRELIMINARIES

**Material, labelling**

The material that was used in this paper is identical to the material already described in ten Bosch (1994). It consists of 191 Dutch, grammatically well-formed utterances (Kraaijeveld, 1994). Of these utterances (elicited speech), 104 were spoken by a male, and 87 by a female speaker. In total, there were 15 female speakers and 18 male speakers.

In table 1, an overview is given of the intonation labels transcribed by intonation experts. The categories '1', '2', 'A', 'B', and 'P' ('pointed hat') constitute over 90 percent of all labels (excluding the 'nulls' which refer to non-labelled syllables).

**Preprocessing**

The pitch contour of the utterances has been processed in a way described in ten Bosch (1994). Firstly, the pitch contour was determined by a subharmonic summation method (Hermes, 1988), followed by a dynamic-programming routine to remove octave jumps. The resulting pitch is transformed to an Equivalent-Rectangular-Bandwidth (ERB) frequency scale since

0-7803-2431-5/95 $4.00 © 1995 IEEE

Table 1: Frequency count of all labels used in the database. The symbol 'P' refers to the rise-fall ('pointed hat') pattern. The category 'other' contains unclear pitch realizations (8), additional labels referring to pitch level ('&', '0') (10), and additional labels denoting a merged category 1&2 (12).

| cat | label | count |
|---|---|---|
| rises | 1 | 187 |
| | 2 | 92 |
| | 3 | 4 |
| | 4 | 3 |
| | 5 | 2 |
| falls | A | 102 |
| | B | 93 |
| | C | 6 |
| | D | 28 |
| | E | 4 |
| rise-fall | P | 113 |
| other pos. cat. | | 30 |
| total pos. cat. | | 664 |
| nulls | | 968 |
| all | total incl. nulls | 1632 |

there are indications that pitch movements lend equal prominence to syllables when expressed on this physiological frequency scale (Hermes & van Gestel, 1991). Finally, in order to allow a comparison across speakers and utterances, the pitch contours have been normalized for each utterance by subtracting the mean pitch, and subsequently dividing the transformed pitch contour by its variance.

## 3. CLASS SEPARABILITY (step A).

In ten Bosch (forthcoming), it has been shown that the classification of the intonation category depends on four choices in the classification model that are all of substantial importance:
(1) the choice of the input representation, i.e. the step from the actual pitch contour to a (vector) representation of that contour suitable as input for the classifier.
(2) The classification method itself is based on the Multi-Layer Perceptron (MLP). For the sake of comparison, the Bayes-Gauss-Luce (BGL) model (an extension of linear discriminant analysis) is considered as well.
(3) The error criterion.
(4) The intonation grammar. The grammar specifies which sequences of categories are feasible given an utterance.

It appears that these four issues are important for the classification results as well as for the underlying interpretation of the classification. With respect to the input representation, it has been found that a three-syllable window performs adequately, all other methodological choices being kept fixed. Vectorial representations in which some of the components were based on pitch contour information from the neighbouring syllables leads to significantly better results than in the case of a two- or one-syllable representation.
Comparing BGL to MLP, it has been shown that the MLP with one hidden layer performs significantly better than does BGL, as soon as the hidden layer consists of sufficiently many hidden units (from 4 on). An MLP with hidden layer performs significantly better than does an MLP without a hidden layer.
The error criterion has been chosen in order to yield an interpretation in line with the Viterbi backtracking algorithm. The following similarity measure was chosen (to be optimized) (cf. Fukunaga, 1972; ten Bosch & Smits, forthcoming):

$$\sum_X \sum_i y_i \tilde{y}_i$$

the first and second summation denoting a sum over all stimuli and over all categories, respectively. Here, $y_i = P_{observed}(C|X)$ and $\tilde{y}_i = P_{predicted}(C|X)$ denote the observed and predicted categorical probabilities given a certain stimulus (representation of a pitch movement) $X$. The output of the MLP was normalized in order to comply with the normalization constraint $sum_C P_{predicted}(C|X) = 1$ for all $X$.
The first three issues refer to step (A) of the algorithm. In tabel 2, the classification results are shown for the BGL-method as well as for the best MLP-topology. The results are based on cross-validated models. The topology of the MLP in this case was 4-10-6, referring to a four-dimensional input representation, 10 hidden units, and 6 output 'response' categories (i.e. the six categories mentioned in the first column).

## 4. GRAMMAR (step B).

The intonation grammar as presented in HCC is put forward as being generally suitable for the description the intonation of well-formed Dutch utterances. In order to render this grammar applicable to the present purpose, it was pruned so as to contain only the categories '1', '2', 'A', 'B', and 'P'. Accordingly, the only allowed expansions in case of e.g. the '1A' pattern for a sentence of three syllables are 'null-1-A', '1-null-A', and '1-A-null'. Theoretically there are infinitely many of the 'basic patterns' such as '1A', which refer to the compact description of the intonation contour of an utterance without

Table 2: Confusion matrices (presenting absolute occurrences) generated by the classification based on Bayes-Gauss-Luce (upper panel), and on the 4-10-6-MLP (lower panel). This MLP gave the best results on independent test sets. The stimuli are arranged along the columns; responses are given along the rows. Each category was represented by 67 different acoustic realizations.

| | stimulus S | | | | | |
|---|---|---|---|---|---|---|
| resp. R | null | '1' | '2' | 'A' | 'B' | 'P' |
| '-' | 24 | 3 | 11 | 8 | 15 | 3 |
| '1' | 4 | 46 | 5 | 0 | 0 | 14 |
| '2' | 11 | 7 | 36 | 1 | 8 | 7 |
| 'A' | 9 | 0 | 1 | 47 | 12 | 1 |
| 'B' | 16 | 1 | 6 | 11 | 29 | 3 |
| 'P' | 3 | 10 | 8 | 0 | 3 | 39 |
| '-' | 38 | 0 | 5 | 1 | 4 | 1 |
| '1' | 2 | 59 | 2 | 0 | 0 | 5 |
| '2' | 2 | 5 | 55 | 0 | 3 | 7 |
| 'A' | 9 | 0 | 0 | 62 | 8 | 0 |
| 'B' | 14 | 0 | 2 | 4 | 50 | 2 |
| 'P' | 2 | 3 | 3 | 0 | 2 | 52 |

specification of the location of the pitch movements. In practice, their number could be reduced to nine. These basic patterns are listed in table 3. The third column in this table refers to a 'complexity' measure, which is in this case defined to be the number of terminal nodes ('labels') in the grammatical pattern.

In table 4, the results of the entire algorithm are shown ordered by the basic grammatical pattern (column 1). The second and third columns contain the absolute and relative number of occurrences of the patterns in the database, respectively. The classification result is considered in two ways. The weak version (columns 4 and 5) refers to the weak error comparison, in which the human categorical pattern was considered to match perfectly with the best 10 predicted patterns as soon as the patterns were partially correct. For example, the human transcription was 'PP1A' and the best model prediction was '1A1A' and the second 'PP' with the P's at the correct location. In that case, the weak match is perfect. In the strong version (columns 6 and 7), the match is perfect only if the best predicted pattern coincides with the human transcription.

Table 3: Number of terminal leaves as a measure of the complexity of each grammatical pattern.

| index | gram. pattern | complexity |
|---|---|---|
| 1 | (1A) | 2 |
| 2 | (1A)(1A) | 4 |
| 3 | (1A)B(1A) | 5 |
| 4 | P | 1 |
| 5 | P(1A) | 3 |
| 6 | PP | 2 |
| 7 | P2(1A) | 4 |
| 8 | P2(12)1A | 6 |
| 9 | P2(12)(12)1A | 8 |

The data in table 4 are obtained for the most minimal grammar out of these nine basic patterns fitting the utterance transcription. As soon as the grammar is too 'sloppy', the error rate increases. This is not shown in the table; in figure 1, however, relevant data are adduced. In this figure, the classification results are shown versus the complexity of the basic patterns as defined in table 3. The circles refer to the weak match, the '+'-signs to the strong version of the match; both refer to the minimal grammar. The cross signs refer to the maximal grammar, i.e. the grammar that allows each of the nine basic patterns, in the weak match mode. Especially for the weak match, there is a gradual trade-off between the complexity of the grammar and the classification result, which can qualitatively be explained by the fact that the search space size increases (worse than polynomially) with the number of terminal leaves, and the decrease of the probability of the correct match. The results in case of the strong match are more difficult to interpret, but some of the data points can be traced back to the fact that the *a priori* biases for the basic patterns are not equal.

So far we have not been able to detect a consistent relation between (1) the complexity of the grammar, (2) the 'sloppiness' of the grammar, and (3) the classification result.

## 5. CONCLUSIONS

An algorithm aiming at the automatic transcription of pitch contours has been discussed. It consisted of two steps: (A) a syllable-based separation algorithm (a non-linear classifier, modelled by a Multi-Layer Perceptron). In this step, about 80 percent of realizations in the the six categories '1', '2', 'A', 'B', and 'null' can be correctly classified. In step (B), a Viterbi backtracking algorithm is used to extend the syllable-based separation results to sentence-level. Depending on the type of grammar used and especially on its 'sloppiness', the performance of the algorithm varies between 50 and 94 percent in the 'weak comparison case' with minimal grammar. In case of the

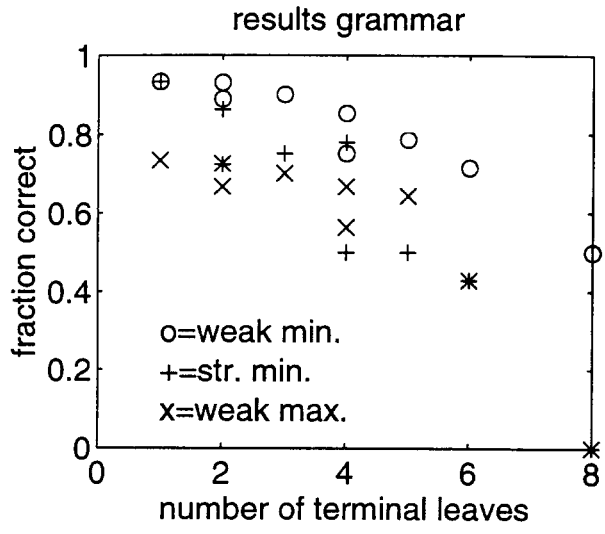


Figure 1: Classification results versus the complexity of the grammar in three cases. For a description see the text.

Table 4: Results after invoking the minimal grammar. In the first column, the grammatical pattern is shown. The absolute and relative occurrences in the database are shown in the second and third column. For a further explanation see the text.

| | occurrence | | class. result | | | |
|---|---|---|---|---|---|---|
| gram. pattern | abs. | rel. | weak | frac. | str. | frac. |
| 1A | 29 | 0.18 | 27 | 0.93 | 25 | 0.86 |
| 1A1A | 27 | 0.17 | 23 | 0.85 | 21 | 0.78 |
| 1AB1A | 14 | 0.08 | 11 | 0.79 | 7 | 0.50 |
| P | 15 | 0.09 | 14 | 0.93 | 14 | 0.93 |
| P1A | 20 | 0.13 | 18 | 0.90 | 15 | 0.75 |
| PP | 18 | 0.11 | 16 | 0.89 | 13 | 0.72 |
| P21A | 16 | 0.10 | 12 | 0.75 | 8 | 0.50 |
| P2121A | 14 | 0.08 | 10 | 0.71 | 6 | 0.43 |
| P212121A | 6 | 0.04 | 3 | 0.50 | 0 | 0.00 |
| total | 159 | 1.00 | 134 | 0.84 | 109 | 0.69 |
| other | 32 | – | – | – | – | – |

maximal grammar, results are generally worse, even in the weak match mode.

# References


[1] Bagshaw, P.C. (1993). 'An investigation of acoustic events related to sentential stress and pitch accents, in English.' Speech Communication, 13. pp. 333–342.

[2] Bosch, L.F.M. ten (1994). 'On the representation of acoustic realizations of Dutch pitch movements.' To be submitted to J. Acoust. Soc. Amer.

[3] Bosch, L.F.M. ten (forthcoming). 'On the automatic classification of pitch movements.' To be submitted to J. Acoust. Soc. Amer.

[4] Bosch, L.F.M. ten, and Smits, R. (forthcoming). 'On the cost function on fuzzy classification.' Manuscript in preparation.

[5] '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.

[6] Fukunaga, K. (1972) *Introduction to statistical pattern recognition.* New York: Academic Press.

[7] Hermes, D.J. (1990). 'Vowel-onset detection.' J. Acoust. Soc. Am. 87, 866–873.

[8] Hermes, D.J. and Van Gestel, J.C. (1991). 'The frequency scale of intonation.' J. Acoust. Soc. Am. 90, 97 – 102.

[9] Hieronymus, J. (1989). 'Automatic sentential vowel stress labelling.' (Tubach, J.P. and Mariani, J.J., eds.) Proceedings of the first European Conference on Speech Communication and Technology (Eurospeech), Paris. pp. 226-229.

[10] Kraayeveld, J., Rietveld, A.C.M., and Heuven, V.J. van (1993). 'Speaker specificity in prosodic parameters.' In: Proceedings of the ESCA-workshop on prosody. Working Papers 41, Dept. of Linguistics, Lund Univ. pp. 264–267.