PITCH VARIATIONS AND EMOTIONS IN SPEECH

Sylvie Mozziconacci
Institute for Perception Research / IPO, Eindhoven, The Netherlands

ABSTRACT

Speech of a Dutch male professional speaker enacting seven emotions was analysed with respect to pitch. Because observed pitch variations could not easily be captured in a two-component intonation model, the perceptual relevance of the differences observed between natural contours analyzed and contours as described by the model was tested. Results showed that modelling the departures from the model did not result in improved recognition performance.

INTRODUCTION

Utterances expressing different emotions show systematic differences with respect to pitch, temporal properties, and voice quality [1, 2, 3]. Recent work on speech recognition and speech synthesis has made clear the need for insight into the prosodic characteristics associated with different emotions. The work to be reported here is part of a study of these characteristics.

It is not known exactly how the relevant prosodic features should be controlled to get optimal recognition of different emotions in synthetic speech. Hence, acoustic analyses and perceptual evaluations of tentative rules are needed. This involves an adequate representation of the acoustic data.

The research reported here focuses exclusively on the role of pitch. In a previous study [1], tentative rules for synthesizing utterances conveying particular emotions were expressed in terms of a two-component model [4]. In this model, one component represents *pitch register* variations, concerning how high or low the utterance is produced in the speaker's overall range. Register is operationalized by means of a baseline which is anchored in the utterance-final low pitch and which has a certain slope. The other component represents *pitch range* variation, operationalized in terms of the distance between local F0 minima and maxima (i.e., the size of pitch changes). For sake of simplicity, the pitch range has been held constant throughout the utterance in formulating tentative rules, although this is not a necessary assumption within the model.

Comparing the output of the tentative rules to the contours in utterances conveying different emotions produced by a human speaker (these last contours will further be called natural contours), we observed that the rule-based contours differed in several respects from the natural contours, for most emotions. This observation gave rise to the questions to be addressed in the current study:

1. What are the detailed characteristics of pitch contours of utterances expressing different emotions?
2. To what extent does the modelling of these characteristics lead to improved recognition of the emotions?

I. ACOUSTIC ANALYSIS

Materials and method

A male Dutch speaker enacted seven emotions (neutrality, joy, boredom, sadness, anger, fear, and indignation), producing three tokens of each of five sentences for each emotion. The five sentences had previously been found to have neutral semantic content (e.g. *Het is bijna negen uur* "It is almost nine o'clock").

Intonation contours were labelled according to the description by 't Hart, Collier and Cohen [4].

Because natural contours differed substantially from synthetic contours produced by means of rules based on the outcome of preliminary research [1,5], an accurate description of the natural contours was needed. Therefore, pitch was measured at six "anchor points" in each utterance: onset, two peaks (all utterances contained two accented words), two values in the intermediate "valley" (after the first peak and before the second peak), and offset.

Results and discussion

The labelling of intonation contours revealed that the so-called 1&A pattern (pitch rise-and-fall on a single accented syllable) was used for all emotions. Indeed, for most emotions it was the most frequently used pattern. This suggests that the 1&A pattern is appropriate for the expression of all emotions under study. Therefore, in order to exclude variation due to phonological structure, further acoustic analyses will be restricted to utterances realised with the 1&A pattern.

For each emotion a "mean natural contour" was calculated by taking the means for the successive "anchor points". They are shown in figure 1. This figure shows that pitch register and range vary systematically as a function of emotion. Further inspection of figure 1 indicates that there are two major sources of deviation between the natural and rule-based contours:

1. Whereas in the rule-based contours all minima fall on a single declining baseline, the utterance-final low pitch in the natural contours does not simply fall in line with the minima in the earlier part of the utterance. Instead, the offset of pitch can be considerably lower than expected on the basis of the earlier part of the contour.
2. Whereas the size of pitch changes is the same throughout the utterance in rule-based contours, this is not the case in the natural contours. Instead, for contours higher in register, the second pitch accent becomes increasingly larger than the first one.

In fact the only emotions for which the natural contours compare well to the rule-based versions are neutrality and boredom. This is not surprising, since the model was developed on the basis of neutral, uninvolved utterances.

The question arises whether detailed modelling of these departures from the rule-based versions may improve identification accuracy for different emotions.

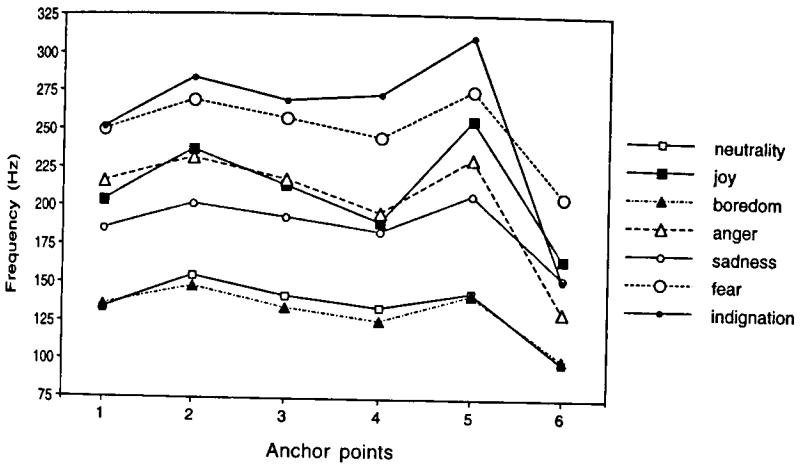

Figure 1. Results of F0 measurements averaged over successive points (so called anchor points) in 1&A realisations of five sentences. Each anchor point is based on at most 15 tokens for each of the seven emotions N: neutrality, J: joy, B: boredom, A: anger, S: sadness, F: fear, and I: indignation. Symbols representing successive anchors for each emotion are connected with lines.

II. PERCEPTUAL EVALUATION

The perceptual test investigates to what extent modelling the detailed aspects of contours for different emotions helps to improve their recognition.

Materials and method

Synthetic pitch contours were generated for a single carrier sentence (*Zijn vrienDIN kwam met het VLIEGtuig* "His girlfriend came by plane"), with accents on /din/ and /vlieg/. Both accents were realized with 1&A type pitch accents.

For each emotion, five different

synthetic pitch contours were produced, representing five conditions. Condition 1 was included for comparison and conditions 2 to 5 resulted in approximations of the pitch contours in figure 1. All synthetic contours had a fixed baseline declination of 3.5 semitones / s.
Sentence duration is 1.77 s.

1. *Two-component optimal perceptual values.*
Contours were generated using values for pitch register and range that had been obtained in a previous adjustment experiment aiming to determine optimal perception-based parameter values [1, 5]. The end frequency and the size of the pitch movements vary as reported in table 1. The slope of the baseline is fixed.

Table 1. Parameter values per emotion for synthetic contours of condition 1. Freq: Endfrequency in Hertz; Exc: Size of the pitch movements in semitones; N: neutrality, J: joy, B: boredom , A: anger, S: sadness, F: fear, and I: indignation.

	N	J	B	A	S	F	I
Freq	65	155	65	110	103	200	170
Exc	5	10	4	10	7	8	10

2. *Two-component best matches to natural contours.*
Values for scaling of the declination line and for the excursion size of the pitch movements were determined to get a close fit to the natural contours (see table 2). The declination line was anchored at utterance onset rather than offset. Table 2 shows end frequencies instead of onset frequencies, to allow comparison with table 1. The distance between the second peak and the baseline (in semitones) was used to determine the excursion size for both pitch accents (this choice was inspired by the fact that the excursion size of the second peak varied much more in relation to emotion than that of the first peak). This means that in this condition the size of pitch movements was equal for both peaks.

Table 2. Parameter values per emotion for synthetic contours of condition 2. Freq: Endfrequency in Hertz; Exc: Size of the pitch movements in semitones; N: neutrality, J: joy, B: boredom , A: anger, S: sadness, F: fear, and I: indignation.

	N	J	B	A	S	F	I
Freq	95	125	100	145	125	160	180
Exc	6	12	5	8	8	9	9

3. *Peak modelling.*
The first peak was manipulated independently of the size of the second one, so as to make the relation between peaks as shown in Figure 1.

4. *Offset modelling.*
Starting from the utterances in condition 2, utterance-final low pitch for each emotion was modelled after the contours in Figure 1.

5. *Peak & offset modelling.*
Condition 5 combines the effects of conditions 3 and 4.

Since for neutrality and joy condition 2 provided accurate offset modelling, contours for conditions 2 and 3 were the same as conditions 4 and 5 respectively. Hence, there were only 31 test utterances instead of 35. A series of 55 stimuli was presented to the subjects; the first 19 gave an idea of the kind and amount of pitch variations allowed in the stimuli, the next 31 were the test-stimuli, and the last 5 were end-of-list fillers.

Sixteen subjects participated in this experiment, which involves a seven-alternative forced choice paradigm with the seven emotion labels. The subjects performed individual interactive listening tests. They listened only once to each stimulus and decided which emotion had been expressed. The 31 test stimuli were presented to different subjects in different random orders.

Results and discussion

Table 3 gives the number of subjects correctly identifying each emotion in the different conditions. Notice that the number of correct responses for neutrality and joy in conditions 2 and 3 are in parentheses. As explained, the contours for neutrality and joy in conditions 2 and 3 had the same utterance-final low pitch as the natural contours so that the contours generated in conditions 2 and 3 for these two emotions actually instantiated conditions 4 and 5. To compare an equal number of judgments for each condition, the results for these stimuli were also included in conditions 2 and 3.

Table 3. Number of correct responses per condition (C1-C5) and per emotion (N: neutrality, J: joy, B: boredom , A: anger, S: sadness, F: fear, and I: indignation.)

	N	J	B	A	S	F	I
C 1	3	3	11	2	2	6	3
C 2	(10)	(7)	6	0	0	5	3
C 3	(9)	(6)	6	1	4	7	3
C 4	10	7	2	0	4	7	9
C 5	9	6	5	0	3	5	5

The total number of correct responses is about the same for all conditions: 30 for condition 1 (Nmax=112), 31 for condition 2, 36 for condition 3, 39 for condition 4, and 33 for condition 5 (chance level = 16). Thus, we find that emotions are recognized better than chance on the basis of pitch alone, but that modelling of contour details does not lead to substantial improvement. Recognition performance in condition 5, which produces the closest match to the natural contours, is similar to the performance in conditions 1 and 2.

The rather low identification performance is probably due to the fact that no characteristics other than pitch have been manipulated. Especially anger and sadness gave poor performance. Previous investigations [1,5] suggested that voice source was an important component for sadness for example. Further looking at the detailed outcome, we observe that there is considerable trade-off between neutrality and boredom. Whereas the condition 1 values for neutrality (judged to be perceptually optimal in a previous study [5]) give a bias towards boredom, the values used for the conditions 2 to 5 give a bias towards neutrality. Further discussion of detailed aspects of the data and of confusions between emotions is beyond the scope of this paper.

CONCLUSION

In sum, we find that, even though there is considerable discrepancy between contours based on a simple two-component model and natural contours, it does not appear necessary to extend the two-component model in order to capture these differences. No clear improvement in recognition is obtained beyond what is achieved in terms of the two-component model.

Furthermore, it is clear that high recognition performance of emotions cannot be obtained through pitch manipulation only, and that other aspects such as duration and voice quality must also be taken into consideration.

ACKNOWLEDGEMENTS

This research was supported by the Co-operation Centre Tilburg and Eindhoven Universities (SOBU).

REFERENCES

[1] Vroomen, J., Collier, R., & Mozziconacci, S. (1993). Duration and intonation in emotional speech. *Eurospeech 93 ESCA Proceedings*, Berlin.

[2] Carlson, R., Granström, B., & Nord L. (1992). Experiments with emotive speech acted utterances and synthetic replicas. *ICSLP 92 Proceedings*, Alberta, 1, 671-674.

[3] Bezooijen, R. A. M. G. van (1984). *Characteristics and recognizability of vocal expressions of emotion*. Dordrecht, The Netherlands: Foris.

[4] Hart, J. 't, Collier, R., & Cohen, A. (1990). *A perceptual study of intonation; an experimental-phonetic approach to speech melody*. Cambridge: Cambridge University Press.

[5] Mozziconacci, S.J.L. (1994). Pitch and duration variations conveying emotions in speech. *IPO Report No. 961*. IPO, Eindhoven.