THE CONTRIBUTION OF PITCH CONTOUR, PHONEME DURATIONS AND SPECTRAL FEATURES TO THE CHARACTER OF SPONTANEOUS AND READ ALOUD SPEECH

Gitta P.M. Laan and Dick R. van Bergem

Institute of Phonetic Sciences, University of Amsterdam,
Herengracht 338, 1016 CG Amsterdam, The Netherlands

ABSTRACT

The separate contribution of the intonation contour, phoneme durations, and spectral features of an utterance to the speech style character was studied by means of a listening experiment. Speech was used from 2 male speakers who each told 'spontaneously' something about themselves and afterwards read out their own transcribed text. Utterances were selected that were identical in wording and that were fluently spoken in both speech styles. The prosodic features pitch, duration, and energy were systematically exchanged between the two speech styles by means of TD-PSOLA. Subjects in the listening experiment were asked to classify the stimuli as either spontaneous or read. It appeared that intonation, phoneme durations, and spectral features all contain cues to a particular speech style, albeit that their separate influence does not dominate over the rest of the information sources of a speech style.

Keywords**:** *Spontaneous speech; read out speech; Speech style perception.*

1. INTRODUCTION

For a perceptual judgement about the speech style, grammar, wording, speech rate, pause structure, and intonation of the spoken tekst are probably mainly responsible. Our project, nevertheless, aims to get an insight in the contribution of spectral and temporal characteristics within smaller speech units like words, syllables, and phonemes in so far as these are relevant to the speech style perception. We will try to relate these spectral and temporal features to functional aspects of speech, such as semantic load and importancce of the word, syllable, or phoneme in its particular context. Furthermore, we would like to determine which spectral and temporal characteristics cause a change in appreciation of speech, as expressed in differences in perceptual judgements on for instance intelligibility and naturalness of speech.

In the present study we first of all wanted to determine to what extent phoneme durations, and spectral features contribute to the character of a speech style. We also studied the effect of the pitch contour in the present experiment, in order to compare its influence on the speech style classification with the influence of phoneme durations and spectral features. A listening experiment was carried out with speech utterances that were identical in wording and grammar for two speech types, to which we will refer as 'spontaneous' and 'read' speech. We systematically exchanged the pitch contour, phoneme durations, and the energy contour between the two speech types, and we also replaced the original pitch contour with a monotonous one. These speech manipulations were done by means of TD-PSOLA [8], which we have experienced to give good quality natural sounding synthesized speech [5,6]. The task of the subjects was to classify the speech style of the stimuli as either spontaneous or read. Depending on the kind of manipulations performed on the speech material, differences in perceptual judgements could be attributed to intonation, temporal, or spectral features of the speech utterances.

This paper only briefly describes the nature of the speech stimuli (2.1), the method of the stimulus manipulations (2.2), and the design of the listening experiment (2.3). Since the listening tests are still going on, only global and preliminary results are given (3). A brief discussion follows in section 4, and in section 5 some preliminary conclusions are given. A detailed description about the experimental procedure can be found in [7].

2. METHOD

2.1 Speech material

Speech stimuli were chosen from the recordings of 2 male speakers (S_1 and S_2) who each told an interviewer 'spontaneously' something about themselves and afterwards read out their own literally transcribed (although somewhat polished for false starts, etc.) spontaneously spoken text [1,4]. From these recordings, pairs of spontaneous and read speech utterances that were identical in wording were selected. These utterances had to be fluently spoken. After this preselection 20 utterance pairs from the speech of S_1 and 22 utterance pairs from the speech of S_2 remained.

The produced speech style according to a strict definition can easaly conflict with the speech style that is actually perceived. However, in order to get a good insight in the contribution of pitch contour, phoneme durations, and spectral features to the character of a speech style, it is necessary to start from utterances that are consistently classified as either spontaneous or read speech. Such clearly distinguishable utterance pairs were selected by means of a listening test in which 17 subjects were simply asked to label an utterance as spontaneous or read. A response was defined as correct if it agreed with the spontaneous or read out style used by the speaker.

According to a sign test, an utterance was correctly classified at a level above chance if at least 12 out of the 17 subjects had a correct response ($p < 0.1$). If this occurred for both the spontaneous and the read version of an utterance pair, we kept them in the selection. This criterion resulted in 8 suitable pairs for S_1, and 11 suitable pairs for S_2.
The small number of utterances that remained, and the average score of only 84.7 % correctly labeled stimuli for these 'well classified' utterances illustrate that spontaneous and read speech can easily blend into each other.

2.2 Preparation of the test conditions

In order to determine the individual contribution of pitch contour, phoneme durations, and spectral features of a speech signal to its speech style character, speech stimuli were created by systematically manipulating these aspects in the selected speech utterances. The original utterances formed a control condition. Altogether the 38 selected utterances ((8 + 11) utterances × 2 speech styles) occurred in the following 5 test conditions:

1. Original speech utterances as control condition (CON).
2. Phoneme durations copied from spontaneous utterances to their read counterparts and vice versa (DUR).
3. Pitch contour replaced by a constant F_0 that was the average of the F_0 medians from the two realisations of an utterance pair (MON).
4. Pitch contour copied from spontaneous utterances to their read counterparts and vice versa (PIT).
5. Phoneme durations, pitch contour, and relative energy of phonemes (total prosody) copied all together from spontaneous utterances to their read counterparts and vice versa (PROS).

For each manipulation, features from the utterance in one speech style (henceforth the *supplying* utterance) were copied to the corresponding utterance in the opposite speech style (henceforth the *receiving* utterance). The receiving utterance was thus always the foundation for the manipulations. Table 1 outlines for each test condition which features in the manipulated utterance originated from the receiving utterance and which features originated from the supplying utterance. Note that in condition DUR the *exchanged* phoneme durations, and in condition PIT the *exchanged* pitch contour, can be seen as the subject of the test, and that in condition PROS the *original* spectral features of the receiving utterance can be seen as the subject of the test.
All 38 speech utterances were lowpass filtered at 4.5 kHz and digitized at a sample rate of 20 kHz with a 12-bit precision. By means of our program Psola (using a TD-PSOLA technique), pitch, duration, and energy of a speech signal can be manipulated individually or in all possible combinations based on pitch markers and a voiced-voiceless labelling of the speech signal. The pitch analysis was done on each speech utterance by means of the program Toon [2]. A phoneme segmentation was done 'by hand' on each speech utterance. After the manipulations with Psola all stimuli were scaled to a fixed overall amplitude level to avoid annoying loudness differences during the listening test.

Table 1. For each condition the contribution of features from the receiving utterance and the supplying utterance to the manipulated utterance is given. (S=spectral; P=pitch; D=duration; E=energy, MP=median pitch).

Condition	Features from receiving utterance	Features from supplying utterance
CON	S, P, D, E	None
DUR	S, P, E	D
MON	S, MP, D, E	MP
PIT	S, D, E	P
PROS	S	P, D, E

2.3 Listening experiment

Altogether, 80 stimuli of S_1 (8 utterances × 2 speech styles × 5 test conditions) and 110 stimuli of S_2 (11 utterances × 2 speech styles × 5 test conditions) were presented. The stimuli were blocked per speaker. Within each speaker block the stimuli had a random order with the restriction that between two versions of the same speech utterance at least three different utterances were presented. Each subject was given a different random order. The order of the 2 speaker blocks was systematically alternated. The two final stimuli of a speaker block also served as 'practice' items at the beginning of the speaker block.
We plan to use 30 subjects for the listening test, but so far the test has only been done by 11 subjects. The subjects were instructed that during the test they would hear several versions of the same speech utterance, but that each time the way in which it was spoken would differ. Their task was to decide whether the stimulus they heard resembled more a spontaneous speech style or a read speech style. Then they had to indicate on a five-point scale how difficult it was for them to choose the speech style.
The listening test was done on line at a computer terminal. The stimulus sentence was written to the screen 300 ms before the acoustic stimulus was presented. In this way listeners knew beforehand *what* was being said, so that they could concentrate on *how* the speech utterance was said. The subjects heard each stimulus twice in succession with an inter-stimulus time of 300 ms. Responses were given by clicking with the mouse in answer blocks of their choice on the screen. The answer blocks 'spontaneous' and 'read', as well as the extremes of the five-point scale 'difficult' and 'easy' were per subject alternately put on the left or the right of the answering screen. After responding on the five-point scale, the next stimulus was automatically presented. In this way each subject could do the test in his own pace.

3. RESULTS

Since the listening tests are still going on, we can only present some global and preliminary results of the listening experiment. For the speech style classification a response on a manipulated utterance was defined to be 'correct' if it agreed with the speech style of the receiving utterance (see Table 1). For instance, the response on a read utterance that was manipulated for any of the conditions DUR, MON, PIT, and PROS, should be labelled 'read' in order to be considered correct. Fig. 1 shows the average percentages correct classification per test condition as scored by the 11 subjects.

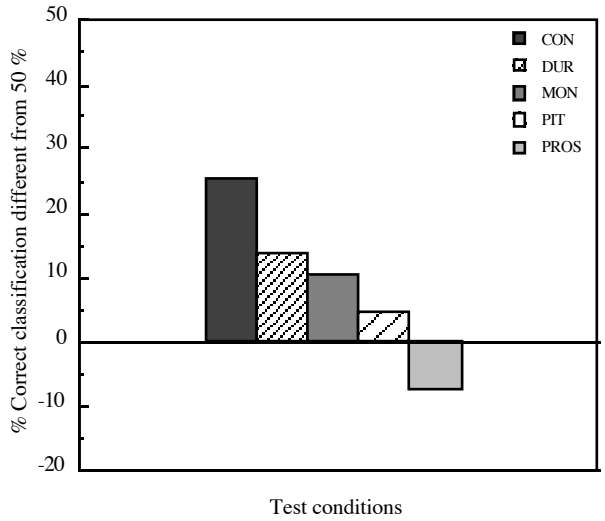

Figure 1. Percentages correct classification per test condition, averaged over all responses to all stimuli from both speakers in both speech styles presented as differences from 50 % which is the chance level.

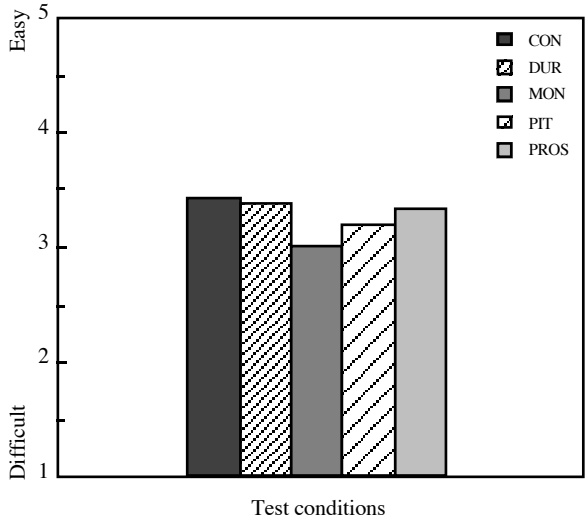

Figure 2. The ease-of-classification scores per test condition, averaged over all responses to all stimuli from both speakers in both speech styles scored on a five-point scale.

The percentages are presented as differences from 50 %, which is the chance level. According to a sign test, a score of 3.9 % above or below 50 % would be above chance level ($p < 0.05$). This margin is surpassed in all conditions.

For the original speech utterances (condition CON) a 75.4 % correct classification score was reached, which is remarkably lower than the 84.7 % correct that was obtained for the 'utterance preselection test' (see 2.1). This is probably due to the fact that in the selection test the stimuli were presented as a blocked condition, whereas in the actual test all test conditions were mixed.

Every manipulation of the speech utterance had a clear influence on the classification of the speech style. The exchange of phoneme durations (condition DUR), caused an 11.5 % decrease in correct classification with respect to the control condition. When the subjects were not given information about pitch movements (condition MON), the classification score decreased with 14.9 % relative to the control condition. If the pitch contour was copied from one speech style to the other (condition PIT), the subjects were even more confused about the speech style, and the percentages correct classification decreased with 20.6 %. However, on the whole the speech style of the three conditions DUR, MON, and PIT was still more or less perceived as the one from the receiving utterance. The total exchange of the pitch contour, phoneme durations, and energy from utterances with their counterparts (condition PROS), on the other hand, did turn round the judgement of the speech style. The classification score diminished with 32.6 % with respect to the condition CON to 7.2 % below the 50 % level.

For each listener the percentage correct responses was calculated per speaker, speech style, and test condition. Each percentage was thus based on 8 utterances from S_1, or 11 utterances from S_2. On these percentages an analysis of variance with repeated measures was done with 3 trial factors: 'speaker' (2 levels), 'speech style' (2 levels), and 'test condition' (5 levels). An inverse sine transformation was applied to the percentages, which gave a better distributed input to the analysis of variance [3]. It appeared that the factors 'speaker' and 'speech style' were not significant ($p > 0.05$), but the factor 'test condition' was significant ($F = 16.0$, $p < 0.001$). There was also a significant interaction effect of 'speaker' with 'speech style' ($F = 13.7$, $p < 0.01$). A test for pairwise contrasts showed that only the conditions DUR, MON, and PIT were not significantly different from each other ($p > 0.05$) for this limited number of listeners.

Fig. 2 shows for each test condition the 'ease-of-classification' that was scored on a five-point scale by the 11 subjects. A score of 1 corresponds with 'difficult to classify', and a score of 5 corresponds with 'easy to classify'.

It can be seen that the subjects experienced the test on the whole as slightly more 'easy' than 'difficult' (all average scores for ease-of-classification were higher than 3.0). There was not any condition that showed a strong preference for one side of the scale. The control condition received the highest score (3.4) for ease-of-classification. For the unnatural sounding utterances from condition MON the score was lowest (3.0). The scores for ease-of-classification in the remaining conditions decreased in the order DUR, PROS, and PIT.

For each listener the average ease-of-classification score was calculated per speaker, speech style, and test condition. Each score was thus again based on 8 utterances from S_1, or 11 utterances from S_2. An analysis of variance was done with the same design as for the classification results. It appeared that the factor 'speaker' and 'speech style' were again not significant ($p > 0.05$), but the factor 'test condition' was significant ($F = 5.6$, $p < 0.001$). There were no significant interaction effects found ($p > 0.05$). A test for pairwise contrasts showed that only the condition pairs CON-DUR, DUR-PROS, and PROS-PIT were not significantly different from each other ($p > 0.05$).

Although there could not be found a significantly different effect in the classification scores, the exchange of only the pitch contour (condition PIT) hindered the subjects more than the exchange of phoneme durations (condition DUR). Also,

although the classification scores for the monotonous utterances were not significantly worse than the scores for the conditions DUR and PIT, the classification of the speech style of the awkward sounding monotonous utterances was harder to do than the classification of the utterances of any of the other conditions.

4. DISCUSSION

Prosodic manipulations by means of TD-PSOLA, within reasonable bounds, result in high quality natural sounding speech. The stimuli for the present experiment were, on the whole, natural sounding speech utterances of high speech quality, since the required modifications were never substantial. The ease-of-classification scores, which were about the same in each test condition imply that even in the condition PROS, for which the Psola manipulations were the most drastic, the stimuli did not sound too 'synthetic' so that it hindered the listeners in their classification task.

The way in which the utterances were chosen already deprived the stimulus material of cues to a speech style concerning grammar, word choice, pause character, and pause structure. These aspects of speech are probably of major importance to the classification of a speech style. The task of our subjects was thus more difficult than it would have been if they had been confronted with larger randomly chosen speech samples from the two speech styles. However, the present experiment was meant to test the influence of only the pitch contour, phoneme durations, and spectral features.

For the limited number of subjects that did the listening experiment so far, the results indicate that pitch contour, phoneme durations, and spectral features of an utterance all contain cues for the classification of a speech style. Exchanging the pitch contour, or the phoneme durations between spontaneous and read utterances, significantly decreased the number of correct classifications of listeners with respect to the scores on the original utterances, although the original speech style was still dominantly present. By exchanging pitch contour, phoneme durations, and energy all together (the entire prosody) between spontaneous and read utterances, the opposite speech style became dominant according to the judgements of the listeners. If the pitch contour was made monotonous, there was still enough information left for the subjects to recover the original speech style. However, in this condition the stimuli were experienced as more difficult to classify than the other stimuli, as indicated by the relatively low scores on ease-of-classification. The scores on ease-of-classification were in general somewhat lower for pitch manipulations than for duration manipulations.

Which aspects will dominate the classification of the speech style of a particular speech utterance will probably depend on the speaker with his idiosyncratic features, the focus of the listener, and the coincidental (prominent) presence or absence of certain features in the speech utterance. The performance of the listeners in the present experiment varied considerably, probably because they used different strategies to classify the speech style. When the listening experiment has been completed (we plan to use a total of 30 subjects), we want to study the responses of the listeners for each separate speaker, speech style, and speech utterance in more detail. We we are now analysing the stimuli in terms of general pitch height, pitch range, phoneme and syllable durations, energy, and spectral features.

In future research, we want to concentrate more on the way, speech characteristics at the segmental level can influence the appreciation of listeners as expressed for example in the intelligibility, naturalness, vividness, etc. of an utterance.

5. CONCLUSION

Based on the preliminary results of only 11 subjects that participated so far in the present experiment, it seems justified to conclude that:

1. The pitch contour, phoneme durations, and spectral features of an utterance all contain cues for the classification of a speech style, albeit that their separate influence is not dominating over the rest of the information sources of a speech style.
2. The exchange of the entire prosody (pitch, duration, and energy) between spontaneous and read utterances reverses the judgement of the listeners about the speech style of the utterances.
3. When the pitch contour is made monotonous in an utterance, the remaining information still contains enough cues to properly classify the original speech style, although the classification is relatively difficult to do.
4. When phoneme durations conflict with the rest of the speech style information in an utterance, the classification is less difficult than when the pitch contour is the conflicting information.

ACKNOWLEDGMENTS

We would like to thank Louis Pols and Florien Koopmans-van Beinum for their discussions about the experiment and their useful remarks on a previous version of this contribution.

REFERENCES

[1] Bergem, D.R. van (1988): "The first step to a better understanding of vowel reduction", Proc. of the Institute of Phonetic Sciences Amsterdam 12: 61-75.

[2] Bergem, D.R. van (1990): "Pitch period estimation by filtering the fundamental frequency out of the speech waveform", Proc. of the Institute of Phonetic Sciences Amsterdam 14: 17-38.

[3] Kirk, R.E. (1982): Experimental design: Procedures for the behavioural sciences, Wadsworth, Inc., Belmont.

[4] Koopmans-van Beinum, F.J. (1990): "Spectro-temporal reduction and expansion in spontaneous speech and read text: The role of focus words", Proc. 1990 International Conference of Spoken Language Processing, Kobe, Japan, Vol. 1: 21-24.

[5] Laan, G.P.M. (1991): "The importance of spectral quality of vowels for speech perception", Report of the Institute of Phonetic Sciences Amsterdam 116, 34 pp. (in Dutch).

[6] Laan, G.P.M., Bergem, D.R. van & Koopmans-van Beinum, F.J. (1991): "The importance of spectral quality of vowels for the intelligibility of sentences", Proc. Eurospeech 91 (3), Genova: 1129-1132.

[7] Laan, G.P.M. (1992): "Perceptual differences between spontaneous and read aloud speech", Proc. of the Institute of Phonetic Sciences Amsterdam 16: 65-79.

[8] Moulines, E. & Charpentier, F. (1990): "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones", Speech Communication 9: 453-467.