Institute of Phonetic Sciences,
University of Amsterdam,
Proceedings 16 (1992), 65-79.

PERCEPTUAL DIFFERENCES BETWEEN SPONTANEOUS AND READ ALOUD SPEECH

Gitta P.M. Laan

Abstract

In this study we tried to find out to what extent the intonation contour, phoneme durations, and spectral features of an utterance contribute to the character of a speech style. For this purpose a listening experiment was carried out with speech utterances from 2 male speakers who each told 'spontaneously' something about themselves and afterwards read out their own literally transcribed spontaneously spoken text. We selected utterances from both speech styles that were identical in wording and that were fluently spoken. By means of TD-PSOLA we systematically exchanged the prosodic features pitch, duration, and energy between the two speech styles. The task of the subjects in the listening experiment was to classify the speech style of the stimuli as either spontaneous or read. It appeared that the subjects could not uniformly classify the speech style of the original (unmanipulated) utterances, which indicates that there was no clear dividing line between the 'spontaneous' and 'read' speech. It also 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.

1 Introduction

When listening to the radio, one can try to judge whether a speaker is reading out a text or talking spontaneously. Such a judgement can be based on numerous features of the speech, like grammar, choice of words, hesitations, filled pauses, speech rate, pause structure, intonation, and carefulness of articulation (see e.g. Levin et al., 1982; Barik, 1977; Blaauw, 1992; Koopmans-van Beinum, 1980). Usually grammatical sentences, regular speech rate and pause structure, a formal intonation, and careful articulation, are easily associated with read out text. These features are displayed in speech of, for instance, a radio news reader. Speech with features like false starts of sentences, irregular and sometimes filled pauses, a vivid intonation, and sloppy pronunciation will come across as being spontaneous. Such features can be met in speech of, for instance, a passer-by who is unexpectedly interviewed.

The speech of a news reader and a passer-by are two extremes. In between (and perhaps even beyond) these two extreme examples of speech style there are many gradations of spontaneous and read out speech. Strictly speaking, one can define read out speech as 'pronouncing a written text directly from paper' and spontaneous speech as 'unprepared, on line formulated speech'. However, a good actor is characterised by his spontaneous way of reciting a text. And how spontaneous is in fact the speech of someone who is telling the very same anecdote at every party? There can be an enormous discrepancy between the speech style typification based on

the way the speech is produced, and the way it is perceived. The extent to which a speaker has prepared himself to speak about a particular subject (and in the case of reading out also his familiarity with the text), probably determines the degree in which his speech will be perceived as spontaneous or read. Speech cannot really be characterised in a discrete way as being spontaneous or read. Rather it can be placed along a continuum that has these two typifications as extreme points on the scale.

Because it is difficult to gather good quality recordings of spontaneous speech (laboratory conditions often diminish the spontaneous character of speech), and because the various speech parameters are difficult to control in spontaneous speech, our knowledge of speech production and perception is mainly based upon research done with read out speech, and often even with nonsense words. To what extent this knowledge can be applied to spontaneous speech is not clear. The fact that people label speech as being either spontaneous or read, suggests that relevant differences do exist between various speech samples. Knowledge of these differences could for instance improve the quality and naturalness of speech synthesis. Thus far speech synthesis is judged to be 'monotonous', dull, and difficult to understand. The need to improve the naturalness of speech synthesis, of course, depends on the application. In some man-machine interactions it might be quite desirable that the user realises he is talking to a machine; the unnatural speech of the machine might prompt him to speak clearly and not to expect any intelligence of the machine.

Research into differences between speech styles usually concentrates on global aspects like grammar, word choice, pause structure, articulation rate, or prosody. Although we do believe these aspects are for a large part responsible for the character of speech, our project 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 importance 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 (Moulines and Charpentier, 1990), which we have experienced to give good quality natural sounding synthesized speech (Laan, 1991; Laan et al., 1991). 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.

The nature of the speech stimuli is described in more detail in sections 2.1 and 2.2. The method of the stimulus manipulations is explained in sections 2.3 and 2.4. In section 2.5 the design of the listening experiment is described. Since the listening

tests are still going on, only global and preliminary results are given in section 3. A brief discussion follows in section 4, and in section 5 some preliminary conclusions are given.

2 Method

2.1 Speech material

In order to be able to exchange pitch contour and phoneme durations between the two speech styles, it was essential that the speech material in both speech styles was identical in wording. This condition also ruled out any grammatical, or word choice differences between the two speech styles (in which we were not interested). Speech stimuli were chosen from the recordings of two male speakers 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. The first speaker (S_1) was asked to tell about his favourite dish (Van Bergem, 1988); the second speaker (S_2), a professional news reader, was asked to tell about his career (Koopmans-van Beinum, 1990).

From these recordings, pairs of spontaneous and read stretches of speech that were identical in wording were selected. These stretches had to form an utterance by itself. They also had to be fluently spoken, implying that both realisations of the utterance pairs should be free from filled and unfilled pauses, and from mispronunciations. This selection criterion ruled out global differences in temporal structure and articulation of the utterances (in which we also were not interested). After this preselection 20 utterance pairs from the speech of S_1 and 22 utterance pairs from the speech of S_2 remained.

As was mentioned before, spontaneous talk does not really have to sound spontaneous, and a read out text, on the other hand, can sometimes sound quite spontaneous. 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 subjects were simply asked to label an utterance as spontaneous or read.

The test consisted altogether of 106 stimuli: 84 preselected speech utterances ((20 + 22) utterances × 2 speech types), plus a set of 21 repetitions, plus 1 'practice' stimulus at the beginning of the test. The repetitions merely served to unbalance the presentation of just one read and one spontaneous version of an utterance; they were disregarded in the processing of the results of the test. The stimuli were placed in a random order with the restriction that two different versions of the same utterance should at least be three stimuli apart from each other. The test was done by 17 subjects at a computer terminal. Each subject was given a different random order. The stimuli were presented on line to the subjects. The subjects responded to the stimuli by clicking the answer block of their choice (either 'spontaneous' or 'read') with the mouse. By responding, the next stimulus was automatically elicited. In this way the subjects were able to do the test in their own pace. 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 percentages correct classifications for these particular utterances in the listening test are shown in Table 1 per speaker and per speech style.

Table 1. Percentages correct classification of the speech style on the eventually selected 38 utterances per speaker and per speech style.

Speaker	Speech style	Number	% Correct
S_1	Spontaneous	8	83.1
S_1	Read	8	85.3
S_2	Spontaneous	11	86.1
S_2	Read	11	84.0
S_1 and S_2	Spontaneous and read	38	84.7

The small number of utterances that remained, and the average score of only 84.7 % for these 'well classified' utterances illustrate that spontaneous and read speech can easily blend into each other. Table 2 lists the 19 Dutch speech utterances together with an English translation. These utterances were the ones actually used in the present experiment.

Table 2. The upper section lists the 8 speech utterances from S_1, and the bottom section lists the 11 utterances from S_2 that were used in the present experiment.

Speech utterance	English translation
...want die worden gewoon gekweekt...	...for they are just breeded...
...zodat ze lekker bruin worden...	...so that they become nicely browned...
...ze zijn eigenlijk heel snel klaar...	...they actually are very quickly done...
...ik eet het wel eens met Mieke...	...I sometimes eat it with Mieke...
...dan zet ik het vuur nog even extra hoog...	...then I turn up the fire extra for a while...
...en dan gaat het plezier er eigenlijk al af...	...and then the pleasure is already more or less gone...
...de bereiding is eigenlijk ontzettend simpel...	...the preparation is actually very simple...
...en dat je niet continu met je vingers in je mond zit...	...and you're not continuously with your fingers in your mouth...
...die man had een confectiefabriek...	...that man owned a clothing factory...
...dat nieuws werd in Nederland gehoord...	...that news was heard in the Netherlands...
...daar ben ik veertig jaar gebleven...	...there I stayed for forty years...
...dat was een kwestie van financiën....	...that was a matter of finance...
...dat was een extra schnabbeltje erbij...	...that was an extra job on the side...
...dus ik had daar de grootste moeite mee...	...so I had the greatest difficulty with that...
...zo heb ik in de praktijk het vak geleerd...	...that's how I learned the job in practice...
...na verloop van een half jaar deed ik het zelf...	...after a period of six months I was doing it myself...
...nou dan heb ik nog een lijstje met vragen hier...	...well, then what's left is a list of questions...
...de sigaretten kostten vijftien cent per pakje toen....	...the price of cigarettes then was 15 cents a package...
...nadat ik veertien dagen ook nog papier heb verkocht...	...after I also have sold paper for a fortnight...

2.2 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.

We expected the pitch contour to have a large influence on the classification. To see how much information about a speech style is left when the pitch contour is lacking, a test condition was constructed in which the pitch contour was monotonous. The pitch contour was replaced by a constant F_0 that was the average of the F_0 medians from the two realisations of an utterance pair. By using this average value a possible effect of global pitch height that might be characteristic for a speech style, was excluded. The other manipulations consisted of copying the pitch contour, or the phoneme durations, or the pitch contour and phoneme durations and energy all together, from a speech utterance in the one speech style to its counterpart in the other speech style. The original utterances formed a control condition. Altogether the 38 selected utterances ((8 + 11) utterances × 2 speech styles) occurred in 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 (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 3 outlines for each test condition which features in the manipulated utterance originated from the receiving utterance and which features originated from the supplying utterance. The resulting stimuli in the conditions DUR, PIT, and PROS consisted of conflicting information about the speech style. In the condition DUR, for instance, the spectral, pitch, and energy features all represent the speech style of the receiving utterance, whereas the duration represents the opposite speech style of 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. A listening test would reveal how confusing these mixtures of speech styles are, and which combinations of features are dominant for the classification of the speech style.

Table 3. For each condition the contribution of features from the receiving utterance and the supplying utterance to the manipulated utterance is given.

Condition	Features from the receiving utterance	Features from the supplying utterance
CON	Spectral, Pitch, Duration, Energy	None
DUR	Spectral, Pitch, Energy	Duration
MON	Spectral, Median pitch, Duration, Energy	Median pitch
PIT	Spectral, Duration, Energy	Pitch
PROS	Spectral	Pitch, Duration, Energy

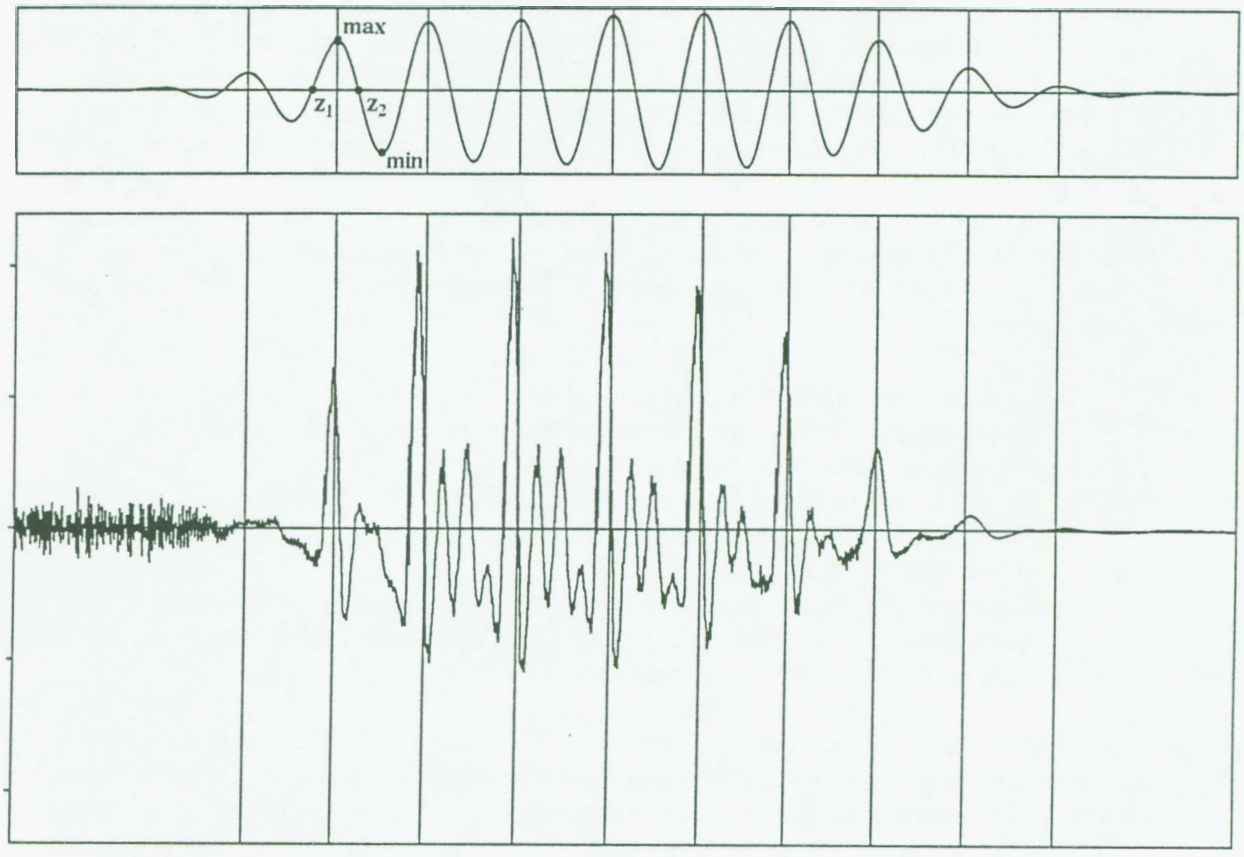

Figure 1. The oscillogram of the vowel [ɪ] from the function word "zit" ([sɪt]) is shown together with the filtered signal that represents the fundamental frequency. The minimum (min), maximum (max), and the two zero-crossings (z_1, z_2) of a period are potential positions for the pitch markers. In this case the energy in the original speech signal is most highly concentrated at position max. Therefore, pitch markers are placed at this point in each period of the vowel.

2.3 Preparation of the speech material

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, prosodic features of speech can be manipulated on the basis of 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 (Van Bergem, 1990). This program filters the fundamental frequency out of the speech signal. This results in a sinusoidal signal in which four points per period (maximum, minimum, and two zero-crossings) are considered as potential positions for pitch markers (see Fig. 1). To model the pitch contour, the actual position of pitch markers is in principle arbitrary, as long as the distances between the subsequent ones accurately reflect the periodicity in the signal. However, it appeared that the speech quality after the manipulations with Psola improved, if pitch markers were placed closer to concentrations of energy in the speech signal. Therefore, the program Toon determines for each separate voiced part in the utterance, at which point in the speech

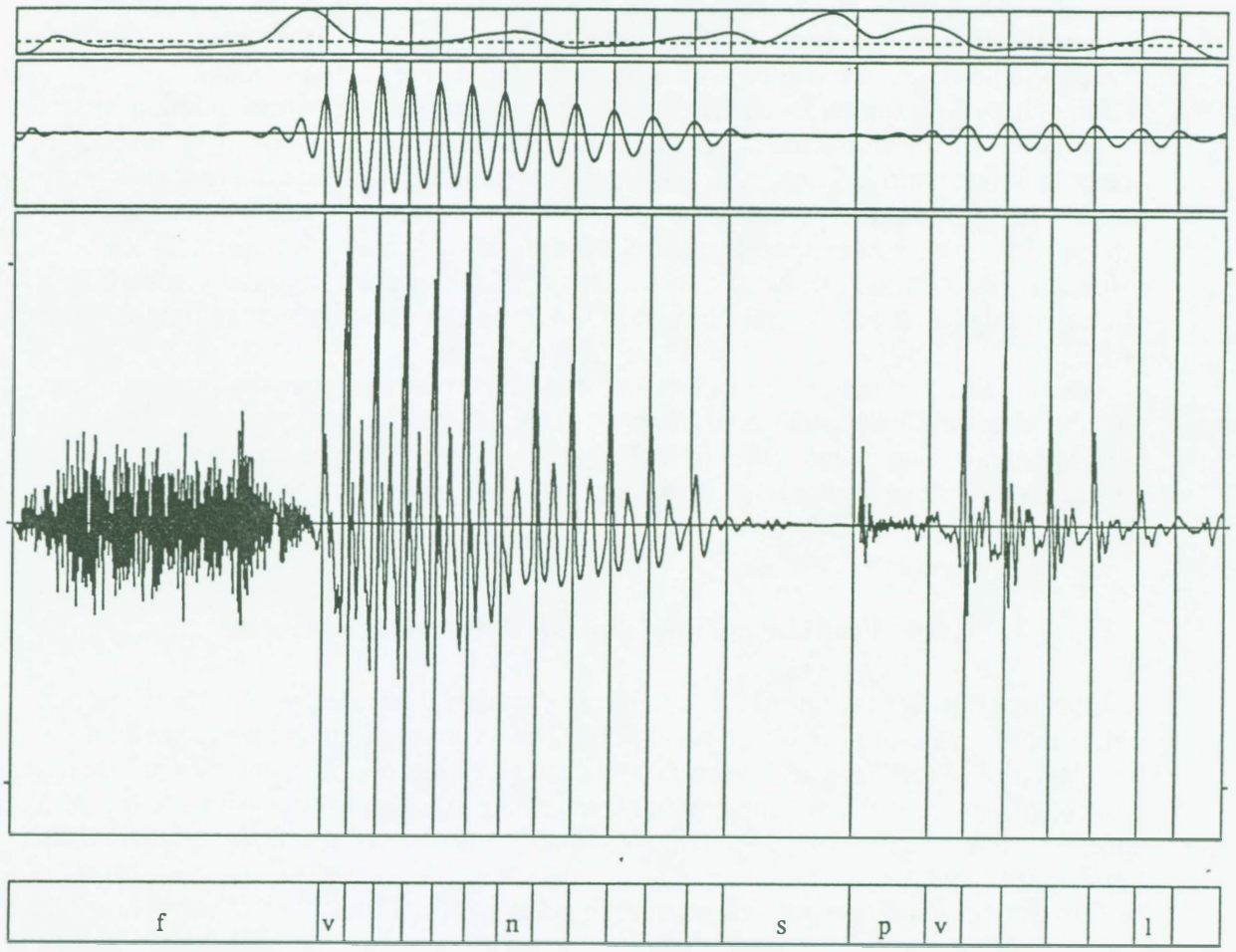

Figure 2. The labelling of the word "simpel" (meaning "simple"). Apart from the oscillogram in the large window, the amount of spectral variation is shown in the top window. The second window from the top shows the filtered signal on which the position of the pitch markers is based. In the bottom window the labels of the phonemes (from left to right: fricative, vowel, nasal, silence, plosive, vowel, liquid) are shown. The phoneme boundary is always directly to the left of the phoneme label.

signal corresponding to one of the four candidate positions for pitch markers, the energy is most highly concentrated. At this point (in Fig. 1, this happens to be the maximum), pitch markers are placed throughout that entire voiced speech part.

After this pitch analysis, a phoneme analysis was done on each speech utterance. This was necessary for a proper exchange of prosodic features per phoneme from one utterance to the other. For this purpose we developed the computer program Label. All phonemes were hand labelled according to the classification: vowel, nasal, liquid, voiced fricative, voiceless fricative, plosive, voice bar, or silence (note that the plosives are divided in a burst, and a silence or voice bar). Phoneme boundaries were determined by ear and by visual inspection of the oscillogram. Also peaks in the 'spectral variation contour' were sometimes helpful in deciding upon the phoneme boundaries. This contour was calculated with a method described in Van Bergem (1993). The phoneme boundaries always coincided with pitch markers, except of course for boundaries between voiceless phonemes, which were added by hand. Fig. 2 gives an example of the labelling of the word "simpel" (meaning "simple"). It shows

the extensive working screen of the program Label. Apart from the oscillogram in the large window, the amount of spectral variation is shown in the top window. The second window from the top shows the filtered signal on which the position of the pitch markers is based. In the bottom window the labels of the phonemes are shown. The phoneme boundary is always directly to the left of the phoneme label.

Especially in a spontaneous speech style a speaker will sometimes delete or insert a phoneme. In the more formal read out speech style, these deletions or insertions occur less frequently. Thus, it occasionally happened that the phonemes that were present in the utterance from the one speech style were not realised in the same utterance from the other speech style. Such a missing phoneme in an utterance with respect to its counterpart was also indicated in the speech signal by inserting a phoneme label that has its start and end marker on the same speech sample (a 'null label').

The positions of the pitch markers that hold information about the pitch contour, and the phoneme labels that hold information about voicing and phoneme durations were written to a so called 'label file'. This label file supplied the program Psola with all necessary information about the speech signal to properly copy prosodic features from one utterance to another.

2.4 Prosodic manipulations with the program Psola

Our program Psola can manipulate pitch, duration, and energy of a speech signal individually or in all possible combinations. The manipulations are performed on the basis of pitch markers and a voiced-voiceless labelling of the speech signal. Psola works with three different methods to manipulate pitch, duration, and energy of a speech signal: one for voiced speech parts, one for unvoiced speech parts, and one for bursts of plosives.

For the manipulations of voiced speech parts by TD-PSOLA (pitch-synchronous overlap-add, in the time domain), building blocks of speech are used that are obtained by multiplying a speech signal with a Hanning window two periods wide and centred at a pitch marker. A 'new' speech signal can be constructed by placing building blocks with their centre pitch markers at certain distances from each other and adding possibly overlapping speech samples. The pitch of the synthesis is controlled by the choice of distances between the centres of the building blocks. The duration of the synthesis is manipulated by repeating or omitting building blocks. The building blocks in the transition parts (if possible, the three initial and final building blocks, depending on the available and desired phoneme duration) are always used and never duplicated in order to preserve as much as possible the spectral dynamics from one phoneme to the next. By scaling each building block with an appropriate factor before constructing the new signal, the energy of the synthesis is manipulated.

Voiceless speech parts are divided in 'noise pieces' of 2.5 ms by the program. The absence of voicing implies that pitch cannot be manipulated. The duration of the synthesis is again controlled by choosing the appropriate number of noise pieces that are to be used for the synthesis. If a noise piece has to be repeated in order to lengthen the speech signal, the samples of the repetition are placed in the reverse order with respect to its preceding use. This is done to avoid periodicity ('buzzing') which will arise if a noise piece is simply repeated several times. Here also, the noise pieces at the beginning and at the end of a voiceless speech part are always used and never duplicated. The energy of voiceless parts is manipulated by multiplying all samples in the entire segment with one appropriate scaling factor.

When a speech part is labelled as a plosive, a synthesis procedure that resembles the one for voiceless speech parts is applied. The difference is that the speech signal is divided in smaller pieces of 0.75 ms which was found to make the synthesis sound more natural. Also, the samples of repeated noise pieces are not placed in reverse order so that the 'bursting character' of the plosive is preserved.

Although spectral characteristics cannot be directly altered by means of TD-PSOLA, the spectral features of a speech signal can in a way be manipulated by the trick of copying the prosodic features (pitch, duration, and energy) of that speech signal to the building blocks of another speech signal that encloses the desired spectral characteristics. The result will be a synthesis that sounds exactly as the original, except for the spectral features that are adopted from the other speech signal. This is in fact done in the condition PROS (see Table 3): The manipulation can be interpreted as 'copying' the spectral features of the receiving utterance to the supplying utterance.

Since all phonemes were labelled in the speech utterances for the experiment, it was possible for Psola to align the phonemes of the two items of an utterance pair, and subsequently copy prosodic features per phoneme. If a phoneme was missing in the supplying utterance (indicated by a null label) with respect to the receiving utterance, the value of the prosodic feature that had to be applied to the phoneme was thus 'nil', and the phoneme was rightly left out in the synthesis. If on the other hand, the phoneme was missing in the receiving utterance (indicated by a null label) with respect to the supplying utterance, Psola could not dispose of building blocks for the synthesis and consequently the phoneme necessarily had to be absent in the synthesis.

After the manipulations with Psola all stimuli were scaled to a fixed amplitude level to avoid annoying loudness differences during the listening test.

2.5 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 15 subjects. These subjects were all students and did not participate in the 'utterance selection test' (see section 2.1). They 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 answering screen is shown in Fig. 3. The stimulus sentence was written to the screen 300 ms before the 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 the answer block of their choice.

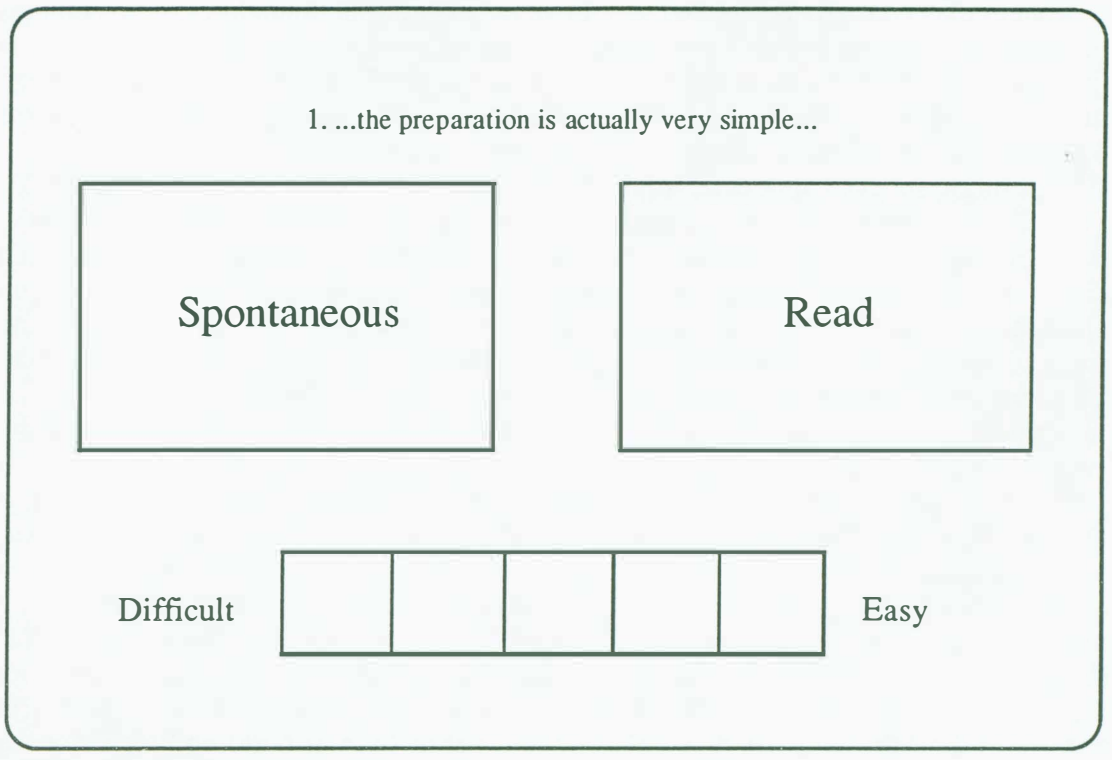

Figure 3. The layout of the answering screen during the listening test.

The answer blocks 'spontaneous' and 'read', as well as the extremes of the five-point scale 'difficult' and 'easy' were for each 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. According to a sign test, a subject would have a classification score above chance level in the control condition if at least 24 out of the 38 utterances were correctly classified ($p < 0.05$). The speech style of all 38 utterances we used in this experiment could on the average be correctly classified 84.7 % of the times in the 'utterance selection test' (see section 2.3). Nevertheless, 4 subjects out of the 15 that participated in the actual experiment did not reach a classification score above chance level in the control condition. The scores of these subject were therefore not used in the processing of the results.

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 3). 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. 4 shows the average percentages correct classification per test condition as scored by the (remaining) 11 subjects. 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.

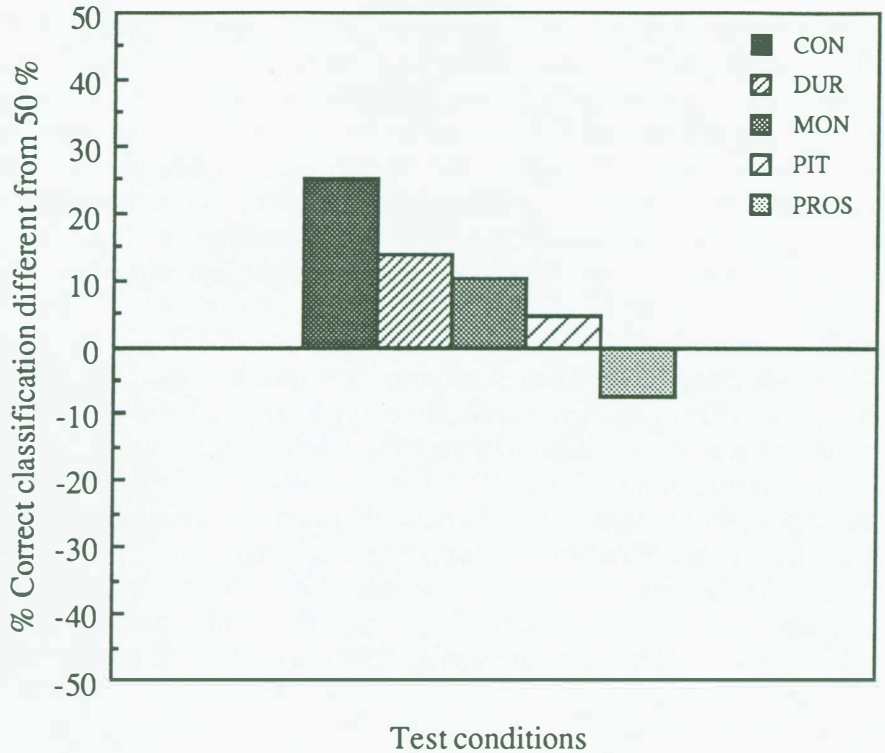

Figure 4. Percentages correct classification per test condition presented as differences from 50 % which is the chance level.

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 selection test' (see Table 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 condition CON. 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.

One might expect that the percentages correct scores for the conditions CON and PROS would be more like mirror images with respect to the 50 % level. This difference in deviation from the 50 % level between the condition CON and the condition PROS can be ascribed to the spectral features from the receiving utterance that were still present in the manipulated utterances of the condition PROS (see Table 3). However, apart from the spectral features, also the voice quality of the receiving utterance was still present, because the source characteristics are inherently linked to the spectral features through the synthesis method. Voice quality might sometimes

have supplied cues for a speech style especially in the case of speaker S_2 who was occasionally 'hoarse' in his spontaneous utterances. Also the realised phoneme deletions and insertions sometimes formed excellent cues to a speech style. These factors may all have counteracted the tendency to respond with the opposite speech style in the condition PROS. To really establish the contribution of spectral features to the character of a speech style, it will be necessary to examine the results for only those utterance pairs that had no deleted or inserted phoneme realisations, and that had a similar voice quality. We plan to do this when the listening tests have been completed.

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 (Kirk, 1982). 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 the conditions DUR, MON, and PIT were not significantly different from each other ($p > 0.05$) for this limited number of listeners.

Fig. 5 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'.

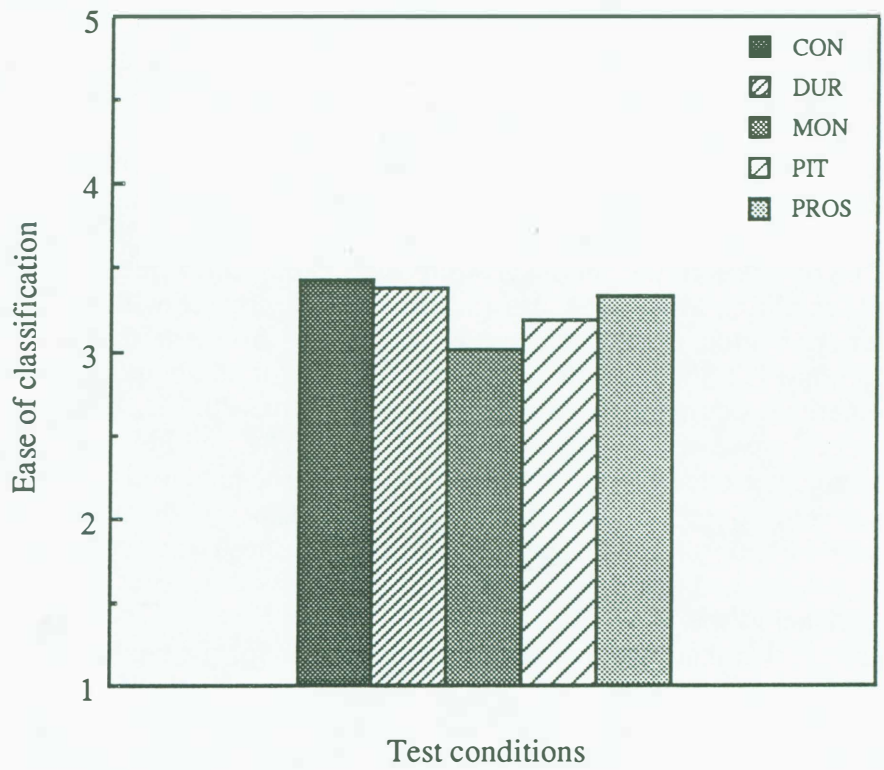

Figure 5. The ease-of-classification per test condition scored on a five-point scale. A score of 1 corresponds with 'difficult to classify', and a score of 5 corresponds with 'easy to classify'.

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 condition CON 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 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. Only an extreme lengthening of the speech signal, or imposing a pitch height that differs too much from the original, so that it conflicts with the spectral quality in the building blocks, gives a somewhat metallic or reverberating synthesis. 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 stimuli in condition MON were of high speech quality, but sounded unnatural because of the missing pitch contour.) Due to a rather large difference in pitch height between the spontaneous speech and read speech of S_2, only utterances from this speaker in the condition PROS (in which both duration and pitch were manipulated) had a noticeably diminished speech quality. Nevertheless, this was not reflected in the ease-of-classification scores, which were about the same in each test condition. This indicates 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. If in future research we will take into account that the global pitch height of the speech material in different speech styles does not differ too much, this will lead to an even higher quality of stimuli.

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. Some subjects were not able to even classify the proper speech style of the original utterances. 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. Furthermore, we will try to relate the scores of the subjects to acoustic measurements of pitch, durations, energy, and spectral features in the speech utterances.

It will be clear that it is necessary to base one's research on many utterances of many speakers in order to get a good insight in the nature of a speech style. The utterances should also preferably be model examples from a certain speech style, for the produced speech style according to a strict definition, can easily conflict with the speech style that is actually perceived. 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. Classifying the speech style of an original (unmanipulated) utterance as either spontaneous or read on the basis of only prosodic and spectral features is not uniformly done by listeners.

2. 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.

3. 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.

4. 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.

5. 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

I would like to thank Louis Pols, Florien Koopmans-van Beinum, and Dick van Bergem for their discussions about the experiment and their useful remarks on a previous version of this article.

References

Barik, H.C. (1977): "Cross-linguistic study of temporal characteristics of different types of speech material", *Language and Speech* **20**: 116-126.

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.

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.

Bergem, D.R. van (1993): "Acoustic vowel reduction as a function of sentence accent, word stress, and word class", to appear in *Speech Communication*.

Blaauw, E. (1992): "Phonetic differences between read and spontaneous speech", *Proc. 1992 International Conference of Spoken Language Processing*, Banff, Canada, Vol. **1**: 751-754.

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

Koopmans-van Beinum, F.J. (1980): *Vowel contrast reduction. An acoustic and perceptual study of Dutch vowels in various speech conditions*, Diss. University of Amsterdam.

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.

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).

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.

Levin, H., Schaffer, C.A. & Snow, C. (1982): "The prosodic and paralinguistic features of reading and telling stories", *Language and Speech*, Vol. **25**, Part 1: 43-54.

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