Expression of emotion and attitude through temporal speech variations

Citation for published version (APA):

Document status and date:
Published: 01/01/2000

Document Version:
Accepted manuscript including changes made at the peer-review stage

Please check the document version of this publication:
• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:
www.tue.nl/taverne

Take down policy
If you believe that this document breaches copyright please contact us at:
openaccess@tue.nl
providing details and we will investigate your claim.

Download date: 22. Oct. 2023
The present study investigates temporal characteristics of speech conveying emotion and attitude. First, a production study was conducted at the global level of the whole utterance, determining the production values of “global speech rate” for each emotion and attitude involved in the study. A perception study was also carried out at this global level, seeking perceptually optimal “global speech rate” values for conveying each emotion and attitude. Perception and production values were compared. Second, as more “local information inside utterances” might be specific to particular emotions or attitudes, the study went a step further with the analysis of “local speech rate” relative to neutrality, considering accented and non-accented speech segments separately. The perceptual relevance of variations in “local speech rate relative to neutrality” was then tested in a perception experiment. Both global and local variations appeared to be relevant for conveying emotions and attitudes in speech.

1. INTRODUCTION

Different types of speech variations such as variations in pitch, intensity, voice quality, speech rate and rhythm contribute to the expression and perception of emotion and attitude in speech. For conciseness sake, the term ‘emotion’ will be used in the present paper as short for both notions of emotion and attitude. Various studies investigated prosodic variations conveying emotion in speech. However, relatively few quantitative studies were concerned with the quantitative temporal characteristics conveying emotion in speech (e.g., van Bezooijen, 1984, Kitahara and Tohkura, 1992). Moreover, the focus was generally restricted to differences in speech rate between emotions at the global level of whole utterances. The aim of the present study is to determine the contribution of temporal variations to conveying emotion in speech in Dutch. Temporal characteristics are described across emotions and across speakers, both at the level of the whole utterance and at the local level of speech segments within utterances. At the global level, variations in overall duration of utterances are expressed as ‘global speech rate’, which means that only variations due to simply stretching and compressing the time scale linearly are taken into account. At the local level, variations in duration of accented and non-accented speech segments considered separately are expressed in ‘local speech rate’. Moreover, the present study combines a perception-oriented and a production-oriented approach. Indeed, the perceptual relevance of the results emerging from the production study needs to be tested in order to model only the temporal variations that are relevant to the communication of emotion in speech.

2. GLOBAL SPEECH RATE

2.1. Analysis of Natural Speech

The speech material used in the present study consisted of 315 utterances resulting of the recordings of three Dutch speakers (two male speakers, MR and RS, and one female speaker, LO). Each of them spoke three times the following five sentences considered to be of semantically neutral content: ‘Zijn vriendin kwam met het vliegtuig’ (His girlfriend came by plane), ‘Jan is naar de kapper geweest’ (John has been to the hairdressers), ‘Het is bijna negen uur’ (It is almost nine o’clock), ‘Ze hebben een nieuwe auto gekocht’ (They have bought a new car), and ‘De lamp staat op het bureau’ (The lamp is on the desk). These sentences were spoken while expressing the seven emotions: neutrality (as a reference to other categories), joy, boredom, anger, sadness, fear and indignation. In order to elicit the emotions, the speakers first uttered sentences of semantic emotional content, intended to evoke emotional situations. Once in that mood, they spoke out the test sentences. The overall duration of each utterance was measured. As the five sentences did not induce the realization of pauses, the notion of ‘global speech rate’ can be defined here in its simplest way, as inversely proportional to the overall utterance duration. This global speech rate is related to neutrality, and thus defined as the ratio of mean overall duration of emotional utterances and mean overall duration of neutral utterances of the same sentence by the same speaker. This means that a speech rate value of ‘0.80’ corresponds to a speech rate reduction in emotional utterances of a specific sentence, resulting in a lengthening of 20% of the emotional speech in comparison with the mean duration of the neutral utterances of this particular speaker. The results are presented in Table 1 for each speaker separately and averaged over the three speakers.
The results show that the global speech rate variations used by the three speakers in conveying specific emotions are often in agreement with each other. For instance, speakers agree on expressing joy with a speech rate very similar to the neutral one, and they all reduce their speech rate while expressing sadness and boredom. On the other hand, in the expression of anger, whereas speakers RS and LO slow down their speech, MR speeds up his speech. For fear and indignation, speaker LO differs from the two other speakers by speaking relatively slower. Although such differences give evidence of various strategies for expressing a single emotion, a certain consistency can also be observed in the type of global temporal variation associated with specific emotions.

2.2. Perception Experiment

In this linear approach, global speech rate is inversely proportional to the overall utterance duration. If this approach is not too global for the expression of emotion in speech, optimal speech rates relative to neutrality may be determined for each emotion. It was already encouraging to notice that in the previous production study, speakers were rather consistent in varying speech rates across emotions. The aim of the present perception experiment is now to seek perceptually based optimal values of global speech rate for conveying emotions in speech. If these values correspond with the values emerging from the production study, we know that they have communicative significance. These values could then be applied in rule-based speech synthesis.

Procedure

In order to copy this time-aligned $F_0$ curve, the optimal Dynamic Time Warping (DTW) path was calculated between the emotional utterances and the neutral utterance of the corresponding sentence, so that the temporal correspondence was preserved. The $F_0$ curve was then copied from the emotional utterance to the neutral one by means of PSOLA. Second, the utterance created this way was made equal in duration to the original emotional one by linear compression or expansion via the PSOLA technique. The resulting utterances, thus, had the same pitch curve as the emotional ones, but voice quality, energy, and all other micro-features of duration, were the same as those of the neutral utterance. Starting from this situation, seven temporal variants were created by compressing/expanding the overall utterance by 70, 80, 90, 100, 110, 120, and 130 percent. For each sentence, the seven variants per emotion served as stimuli.

The versions resulting from the overall speech rate manipulations were organized into two blocks, i.e., one for each sentence. Sentence order was counterbalanced across subjects. In a block, the stimulus order was randomly varied per listener. Ten subjects participated in the listening test. Per emotion, subjects could listen to the seven variants over headphones, as often as they wanted. The task was to choose the three variants that best expressed the given emotion label, and to rank these variants in first, second, and third choice. The rank-order values of the three best variants were transformed into a score in which the very best variant received three points, the second best two, and the third one point. Per sentence, the mean score for each variant was calculated.

Results

The variant that received the highest mean score was considered to be optimal for that particular emotion. The optimal relative sentence duration, with corresponding score is reported in Table 2. For the expression of anger and joy, for instance, a speech rate higher than for neutrality was judged to be appropriate, while a lower speech rate seems to suit the expression of boredom and sadness.

<table>
<thead>
<tr>
<th>Emotion</th>
<th>‘Zij hebben een nieuwe auto gekocht’</th>
<th>‘Zijn vriendin kwam met het vliegtuig’</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Global speech rate</td>
<td>Score</td>
</tr>
<tr>
<td>Neutrality</td>
<td>1.00</td>
<td>2.00</td>
</tr>
<tr>
<td>Joy</td>
<td>1.18</td>
<td>2.30</td>
</tr>
<tr>
<td>Boredom</td>
<td>0.65</td>
<td>1.70</td>
</tr>
<tr>
<td>Anger</td>
<td>1.28</td>
<td>1.40</td>
</tr>
<tr>
<td>Sadness</td>
<td>0.79</td>
<td>1.80</td>
</tr>
<tr>
<td>Fear</td>
<td>1.09</td>
<td>1.50</td>
</tr>
<tr>
<td>Indignation</td>
<td>0.78</td>
<td>1.80</td>
</tr>
</tbody>
</table>

Table 2: Values found to be perceptually optimal for global speech rate relative to neutrality, and mean global speech rate. Note that a score of 1 means on average a third place out of seven variants, while a score of 3, i.e., the maximum score, means the best choice out of the seven variants.
**Discussion**

For all emotions except indignation, the speech rate found to be optimal for each of the two sentences differed by less than 10 percent. A comparison of the results from the present perception experiment with those from the previous production study (see right most column in Tables 1 and 2) shows remarkable similarities. For some emotions, the global speech rate of the production and the perception studies are very similar. Note, however, that in some cases the values found to be optimal in the perception test appear to be more extreme than the values measured in the analysis of speech. In other words, when the speakers realize a reduction of speech rate, the corresponding perception value often suggests an even greater speech rate reduction as being perceptually optimal. This is probably due to the fact that during the perception test, speech rate was the only varied element. Indeed, in order to express oneself in natural speech, one calls on various parameters and their combined effects. It seems reasonable that subjects, who could only rely on speech rate for identifying the emotions expressed in the perception test, have preferred values that are somewhat more extreme than the ones used in natural speech.

In addition, the results of the present study correspond rather well with the findings of related studies. For this comparison, values found in related studies were transformed as well as possible in global relative speech rate values. Clear similarities were found with findings of van Bezooijen (1984). The values she found for joy, anger, and fear are 1.20, 1.16, and 0.99 respectively, and compare quite well with the ones found here to be perceptually optimal. The only difference was that the global speech rate of 1.25 she found for sadness, corresponds to faster speech than in the present study. The main difference with findings of Carlson, Granström and Nord (1992) is the neutral speech, spoken out faster than in all the other emotions in their speech ‘material, which makes it quite difficult to compare global speech rates relative to neutrality. Simply comparing the speech rate values of the other emotions with each other shows that joy and anger are expressed using similar global speech rates, i.e., 0.77, and 0.74, respectively, while sadness is ‘uttered with a lower speech rate than these two emotions, i.e., 0.66. This corresponds fairly well with the findings of the present study. The relative speech rate values proposed by Kitahara and Tokhura (1992), i.e., 1.00 for joy, 1.43 for anger, and 0.87 for sadness are also relatively well in agreement with results of the present study.

It is concluded that the results found in this study are at least in qualitative agreement with some other studies, some of them in other languages. Furthermore, the correspondence of the results from the production study with those from the perception study shows the communicative significance of the values found here.

3. **LOCAL SPEECH RATE**

3.1. **Production study**

As it is now assessed that variations in global speech rate contribute to conveying emotion in speech, it seems well worth considering the way temporal variations are realized within utterances. The expression of emotion might well determine the temporal structure of speech at local level. Different types of speech segments could be considered in order to investigate whether a non-linear time distribution occurs in emotional speech. Potential candidates for such a study are, for instance, the rate of final lengthening, the proportion of silent segments in the utterances, the way temporal variations affect the different phonemes (e.g., assimilation or reduction), or the duration of different phonemes. However, the simplest solution, allowing the study of local temporal variations without involving too many details, is to consider the relative length of accented and non-accented speech segments. An accented speech segment is composed of one lexically stressed syllable on which a pitch accent is realized, whereas an unaccented speech segment is composed of one or several syllables in succession, on which no pitch accent was realized. In order to avoid interference of the effect of final lengthening with the effect of accentuation, it was simply decided not to include the last syllable of the utterances in the investigations.

Note that, in the absence of any emotional context, variations in global speech rate might in themselves lead to changes in the relative duration of accented and unaccented speech segments. In order to distinguish this effect from the effect of emotion, it was necessary to compare the duration of segments of emotional speech to the duration of segments of neutral speech produced at comparable speech rates. As a satisfactory, reliable reference concerning the relative duration of accented and unaccented syllables in neutral Dutch utterances could not be found in the literature, an analysis of neutral speech was first carried out in order to provide a baseline for the present study. The purpose here is to investigate whether the temporal distribution over accented and unaccented syllable types is similar in emotional speech and in neutral speech. If this were the case, a linear model would suffice for the description of the temporal phenomena relevant for conveying emotion in speech. Otherwise, an investigation must be carried out as to whether and how the temporal distribution of both types of syllables follows emotion-specific rules.

**Speech material**

**Emotional Speech**

Two of the five sentences that were used in the previous analysis of global speech rate were considered inadequate for use in the present analysis at local level. One sentence was not included in the study because it induced the realization of a pitch accent in final position, which implies that the effect of final lengthening interferes with the effect of accentuation. A second sentence was excluded of the analysis because the speakers did not always realize one single accent on the same syllable. The three remaining sentences each contained two lexically stressed syllables, both inducing the realization of a pitch accent by the speakers. These lexically stressed syllables are underlined in the sentences listed below, and the accented and unaccented segments are separated by a vertical slash. The three sentences used in the investigation are the following: Sentence 1: ‘Zijn vrien[di] kwam met het [vlieg] tuig’, Sentence 2: ‘Jan[an] is naar de [lat] pper ge[weest]’, and Sentence 3: ‘Het is [bij na] ne[gen] uur’. In total, 182 of the 189 candidate utterances (3 speakers × 7 emotions × 3
sentences × 3 trials) were analyzed. Seven of these utterances were disregarded, either because the occurrence of lengthening after a prosodic boundary would make the comparison with other utterances difficult, or because the speaker was not perfectly fluent or did not produce two accents in the utterance. Since these seven utterances were expressions of five different emotions by the three different speakers, it can be assumed that disregarding these utterances will not significantly have influenced the results.

**Neutral Speech**

The three same sentences were used again in a new recording session of the male speaker MR, who now produced neutral utterances at progressively increasing speech rate. This neutral speech at varying speech rate covered the whole range of speech rate variations realized in emotional speech. The recordings resulted in a total of 171 utterances, namely 57 for each of the three sentences.

**Procedure**

The duration of accented and unaccented speech segments was measured for the 182 emotional utterances and the 171 neutral utterances. Duration was measured from the syllable onset of a speech segment to the syllable onset of the next segment. For each utterance, the durations of the two accented segments were summed. The durations of all unaccented segments, apart from the last syllable, were also summed separately. The ratio of the duration of the accented segments and the total duration of both accented and unaccented segments, was then computed. This durational proportion of accented speech segments will be referred to as ‘proportion accented’. In order to constitute a baseline describing the proportion accented in neutral speech, regression lines were calculated considering the three sentences separately, and fitted through the data points representing the proportion accented of all the utterances of the neutral speech recorded at increasing speech rates. The data concerning the emotional speech are considered relative to the data concerning the neutral speech at variable speech rate. Note that among the 182 emotional utterances, 27 were expressing neutrality.

**Results and discussion**

Per sentence, the proportion accented of all neutral utterances at progressively increasing speech rates, are represented in Figure 1 as the regression lines fitted through the data. For Sentences 2 and 3, considering two regression lines instead of one, clearly increased the explained variation; the root mean square distance between regression line and utterances reduced from .0243 for a single regression line to .0227 for two lines for Sentence 2, and from .0229 to .0173 for Sentence 3 (F₁ = 1.74, F₀.05[56,56] = 1.56, p < .05). For Sentence 1, the root mean square distance of .0174 for one line did not get much better with .0171 for two lines; a single regression line was, therefore, considered to suffice for this sentence. The functions describing the regression lines are the following, where x is the overall utterance duration in seconds and y the proportion accented:

Sentence 1: For \( k \), \[ y = 0.4206 - (0.0388 \times x) \]
Sentence 2: For \( k < 1.83 \), \[ y = 0.3246 + (0.0387 \times x) \]
  else, \( y = 0.4845 - (0.0487 \times x) \)
Sentence 3: For \( k < 1.09 \), \[ y = 0.3012 + (0.1026 \times x) \]
  else, \( y = 0.4906 - (0.0712 \times x) \)

**Figure 1:** Proportion accented speech segments presented per sentence and per speaker.
The proportion accented in the emotional speech of the three speakers is presented in Figure 1 for each sentence separately. A different symbol is used for utterances produced by different speakers. The data points situated above the regression lines indicate a lengthening of accented speech segments relative to the duration of accented segments in neutral speech spoken at the same global speech rate. In Sentence 1, this is for instance the case for the expression of anger, sadness, fear, and indignation. A limitation should be noted, however, in the description of the proportion accented in emotional speech because it was computed on the basis of the neutral speech of a single speaker, MR. Moreover, for Sentence 3, this speaker has produced proportions accented that were somewhat higher in the recordings of emotional speech expressing the reference category neutrality than in the neutral utterances recorded at varying speech rates. This difference in proportion accented between the two recording sessions seems in fact to affect all utterances of this sentence by this speaker, in all emotions.

Although the analysis of speech produced by the three speakers shows the presence of differences in the proportion accented and a slight tendency to stretch accented speech segments relatively more than non-accented ones when the global speech rate is reduced, it does not allow the formulation of hypotheses concerning the exact realization of these variations for the expression of specific emotions. This does also not preclude the possibility that the distribution of time over accented and unaccented speech segments does not play a role in conveying emotion in speech. In order to obtain more information about the communicative significance of the time distribution over accented and unaccented speech segments, a perception experiment will be carried out, which may provide evidence about the perceptual relevance of variations in proportion accented for the expression of emotion in speech.

3.2. Perception Study On Local Speech Rate

The previous production study showed the presence of varying lengths of the accented segments in emotional speech. The aim of this experiment is to investigate whether the proportion accented, nevertheless, is perceptually relevant for the expression of emotion in speech, and, if so, to find, per emotion, the optimal proportion accented.

Speech material

Three neutral utterances by speaker MR, each corresponding to one of the three sentences involved in the previous production study, were manipulated by analysis-resynthesis. Two series of stimuli were generated. Both series include the same six conditions. Series differ in global speech rate, while conditions systematically differ in local speech rate. In the Series 1, the overall duration of all stimuli was kept constant, while in Series 2, the global speech rate varied according to the values previously found to be optimal for each emotion (see right most column in Table 2). In Condition 1, no manipulation of proportion accented was carried out, leaving the time distribution of the accented and the unaccented syllables as in the original neutral utterance. In Condition 2, the proportion accented depended exclusively on global speech rate; it was manipulated according to the functions describing the regression lines represented in Figure 1, and obtained from the analysis of neutral speech at variable speech rate. In Conditions 3, 4, and 5, the proportion accented was manipulated in such a way that the accented speech segments were respectively 20% shorter, 20% longer, and 40% longer than in Condition 2, i.e., than they would be in neutral speech at variable global speech rate. In the generation of all stimuli, the values used for pitch level and pitch range were those found to be optimal for each of the seven emotions in a previous study (Mozziconacci, 1998). All stimuli in Conditions 1 to 5 were generated, across emotions, with the same intonation pattern, i.e., the ‘1&A 1&A’ pattern, which was found to be suitable in the expression of each of the seven emotions in a previous study (Mozziconacci, 1998). Indeed, in situations in which one does not want to introduce variability by using different intonation patterns, a combination of ‘1&A’ configurations of pitch movements was found to be the most suitable one for controlling this variation. However, by stretching the accented syllables, the pitch movement ‘A’ (late prominence-lending fall), becomes more audible, which might influence the perception of emotion. Therefore, in order to separate the effect of the increased proportion accented from the effect of the increased audibility of the fall ‘A’ in the ‘1&A’ configuration of pitch movements, a sixth condition was included. In Condition 6, the proportion accented was not only increased by 40%, as in Condition 5, additionally, the ‘1&A 1&A’ intonation pattern was also replaced by the ‘1B 1B’ intonation pattern. The pitch movement ‘B’ does not lend prominence to the syllable, and was synthesized so that it started at the end of the vowel of the accented syllable. As a result, it was hardly, if at all, audible as a pitch movement. Moreover, the last syllable of the utterances was not affected by the manipulations, “in any condition, because of the possible interference with final lengthening. The last syllable was thus kept as it was” in the original “neutral utterances. All manipulations were based on the PSOLA technique (Moulines & Laroche, 1995). The resulting 252 variants (2 series × 6 conditions × 3 sentences × 7 combinations of pitch level and pitch range) served as stimuli.

Procedure

24 subjects participated in this perception test. Stimuli were organized in 3 blocks, i.e., one per sentence, and sentence order was counterbalanced across subjects. Over headphones, stimuli were presented to each subject in a different random order. After listening once to the utterance, subjects had to choose, among the seven emotion categories proposed to them, the one they thought was conveyed by the utterance.

Results

The identification results were subjected to a categorical, cross-classified loglinear analysis (Fienberg, 1980). The model describing best the data was the one in which the effects of PITCH (the combination of pitch level and pitch range), COND (the conditions), and RESP (the subjects’ responses), as well as the interactions between PITCH and RESP, and between COND and RESP were significant. A cluster analysis allowed the formation of two clusters of each three conditions. The first
cluster groups the first three conditions, corresponding to low proportions of accented speech segments. The three last conditions constitute the cluster of high proportions accented. The fact that Conditions 5 and 6 are members of the same cluster indicates that the effect influencing the subjects’ responses is indeed related to the duration of “the speech segments and not to the audibility of the pitch movements. The results are presented in Table 3 for both clusters. Arrows indicate that the number of responses given by the subjects in the corresponding category is significantly higher (↑) or lower (↓) than the predictions of a loglinear model representing the absence of effect of the conditions on the subjects’ responses. Significant deviations from this model are only obtained for neutrality and indignation, but for the perception of these two categories, temporal variations appear to be very important. The proportion “accented has to be limited in the expression of neutrality. Indeed, an increase of this proportion accentuated has a negative effect on the identification of neutrality. A high proportion “accented, on the other hand, contributes to the perception of indignation.

In order to distinguish the effect of global speech rate from the one of local speech rate, the identification performances were considered as pooled over four types of stimuli. Types 1 and 2 were stimuli generated with constant global speech rate, while global speech rate was varied in stimuli of types 3 and 4. Types 1 and 3 were generated with constant local speech rate, while local speech rate varied in stimuli of types 2 and 4. The percentage correct identification is given in Table 4, pooled over these “different” types of stimuli. Considering the mean percentage “correct identification reported in the rightmost column provides information on the effects of global and local speech rates, independently of each other. As to the effect of global speech rate, comparing the stimuli of type 1 with those of type 3, shows an increase in correct identification of 17%, while comparing the stimuli of type 2 with those of type 4, shows an increase of 16%. As to the effect of local speech rate, comparing type 1 with type 2, shows an increase in correct identification of 8%, while comparing type 3 with type 4, shows an increase in correct identification of 7%. On average, modeling global and local speech rate allows an increase in correct identification of 16.5 and 7.5%, respectively. Moreover, both effects seem to add up, as shows a comparison of type 1 and 4, which suggests independent additive effects of global and local speech rate.

5. CONCLUSION
It was confirmed that the temporal variations at the global level of the whole utterance are of primary importance for conveying emotion in speech, especially for the perception of particular emotions, such as boredom. It does not come as a surprise that the contribution of local temporal variations to conveying emotion in speech, is less important than the effect of variations in global speech rate, or variations in pitch. However, despite the fact that the production study only showed tendencies to vary the temporal structure of speech during the expression of emotion, the communicative function of local speech rate such as the one to convey emotion in speech was established in the perception study. This shows the advantages of supplementing production studies with perception studies. Such an approach does not only show the communicative significance of the variability found in the production studies, it can also reveal, as it was the case in the study of local speech rate, the communicative effect of a variable, even if the statistical significance of its effect was not reached in the production study. Indeed, within the limited set of produced utterances analyzed, other interrelated speech variables can vary. By keeping these variables constant in perception studies, the communicative effect can be shown to be significant.

6. REFERENCES