
User Modeling and User-Adapted Interaction **11:** 297–326, 2001.

Modeling Emotion and Attitude in Speech by Means of Perceptually Based Parameter Values

SYLVIE J. L. MOZZICONACCI*
IPO, Center for User-System Interaction, Eindhoven, The Netherlands

(Received 15 July 2000; in final form 27 January 2001)

Abstract. This study focuses on the perception of emotion and attitude in speech. The ability to identify vocal expressions of emotion and/or attitude in speech material was investigated. Systematic perception experiments were carried out to determine optimal values for the acoustic parameters: pitch level, pitch range and speech rate. Speech was manipulated by varying these parameters around the values found in a selected subset of the speech material which consisted of two sentences spoken by a male speaker expressing seven emotions or attitudes: neutrality, joy, boredom, anger, sadness, fear, and indignation. Listening tests were carried out with this speech material, and optimal values for pitch level, pitch range, and speech rate were derived for the generation of speech expressing emotion or attitude, from a neutral utterance. These values were perceptually tested in re-synthesized speech and in synthetic speech generated from LPC-coded diphones.

Key words: attitude, emotion, experimental phonetics, expression, perception, prosody, speech, speech technology

1. Introduction

Spoken communication involves more than just conveying the literal sense of words and sentences. Prosody can add information, or modify the strictly linguistic content of utterances. Indeed, prosody not only carries information on word stress, phrasing and emphasis, but is additionally thought to be strongly related to speaker specific characteristics, and factors such as the expression of the speaker's emotions and attitudes. Extra-linguistic information, given voluntarily or involuntarily by the speaker, is contained in prosodic features such as intonation, tempo, rhythm, precision of articulation and voice quality. Nevertheless, quantitative details of the correspondence between prosodic features and emotion or attitude, are still poorly understood. Related studies concerned with the vocal expression of emotion have reported qualitative analyses of variations occurring in the expression of emotion in speech. Relatively few studies have tried to quantify the relevant dimensions in the speech signal for the purpose of obtaining parameter values for generating emotion in speech (e.g., Williams & Stevens, 1972; van Bezooijen, 1984; Cahn, 1990;

*Currently at: Phonetics Laboratory, Leiden University, PO Box 9515, 2300 RA Leiden, The Netherlands. e-mail: mozziconacci@hotmail.com

Carlson et al., 1992). Furthermore, quantitative control over relevant prosodic features allowing the expression of emotion and attitude, could prove a powerful means for making synthetic speech sound more natural. Modeling this variability in speech is expected to have an impact on the quality of synthetic speech, and therefore, to enhance the usability of speech technologies. Moreover, modeling the emotional state and attitude of a system user is expected to have incidences on the capacity of computer systems to adapt to the user, and the user's mental state and reactions. An adaptive interactive system should express the right emotion, at the right time, and with the right intensity (Bartneck, in this issue). By including emotions and attitude in the user model of computer systems interacting vocally with the users, one could take the user's mental state into account, and adjust the message contents, and the style of the human-machine interaction to the particular user (Lisetti, 1999; de Rosis & Grasso, in press). In dialogue systems, involving speech recognition and speech synthesis, information concerning the emotion or attitude of the user towards the system and/or the topic can be induced from *what* is said by the user, but also from *how* the user said it. As this information is present in the speech signal, 'emotion recognition' could be performed making use of speech parameters. For an adequate spoken response of the system to the user, emotionally colored speech can be generated using synthesis-by-rule. However, for both recognition and synthesis, a set of such rules needs to be established first, as the effect of prosodic parameters on the vocal expression of emotion and attitude has not yet been well quantified.

The focus of this paper is primarily oriented towards perception of emotion and attitude in speech. Indeed, if we want the results to be usable for the synthesis and the recognition of emotion and/or attitude in speech, it is important to be concerned with the impression produced by the speech on the listener, especially since the computer does not experience any affective state. The present study is largely concerned with the ability of listeners to identify the emotion or attitude intended. It is, in fact, this ability of listeners that we want to model when we try to achieve the modeling of expression of emotion and attitude in speech. Therefore, an attractive strategy for determining a set of parameter values appropriate for conveying emotions and attitudes in speech, is to obtain a model of the listener by means of experimentation. Stimuli produced for use in the perception experiments of the present study were generated in the framework of the IPO approach, which is an experimental-phonetic approach to intonation ('t Hart et al., 1990). Perceptually oriented, it performs a data reduction, using perception as a filter in order to avoid modeling variations that are not relevant to perception, and therefore in the present context not relevant to the communication of emotion or attitude.

An important issue is the one of the categories of emotions and attitudes to be used in such a study. Investigations of speech conveying information on the affective state of the speaker have involved a large variety of expressions, including notions such as emotion, attitude, intention, feeling, and even sentence type. Despite the different lists of emotions that have been proposed by different authors (e.g., Izard, 1977;

Plutchik, 1980; Ekman, 1982; Frijda, 1986), no commonly accepted definition and taxonomy of emotion have emerged. This does not necessarily constitute an insurmountable methodological difficulty, as empirically based lists of relevant notions can be used, either involving notions considered useful in the context of verbal user-system interactions, or based on the ability of subjects to rely on empirically based notions. For the sake of conciseness, and because there is no compelling theoretical base for a distinction between attitudes such as indignation and emotions such as fear, a single term, 'emotion', will be used in the remainder of the text, to refer both to notions of emotion and attitude, without further distinction. For the present study, however, we will rely upon an empirical definition of 'emotion' being a selection of categories that are fairly well identified by listeners among a range of different 'emotions'. Moreover, though neutrality is used as a reference for other categories, it will also be referred to as an 'emotion'.

Establishing a relationship between acoustic and perceptual data, requires a systematic approach. Speech material needs to be gathered that can be considered to be representative of the emotions included in the study. The first next step is to examine the correspondence between the perception of emotion and the acoustic correlates of the perception of emotion. Indeed, initial values for the generation of speech samples to be used in perception tests, are inspired by an analysis of speech produced while expressing emotion vocally. Consequently, the procedure adopted here successively involves natural (i.e., human) speech, manipulations of natural speech via analysis-resynthesis, and synthetic speech. Considering the fundamental frequency (F_0), an advantage of this systematic procedure is that the time course of the pitch, as in natural speech, can be approximated by so-called close-copy stylizations (De Pijper, 1983), i.e., F_0 curves (see Figure 1) consisting of straight-line approximations with as small a number of straight-line segments as possible. The minimum number of line segments is just enough to make the F_0 curve, when used in re-synthesized speech, perceptually identical to the original F_0 curve. By using such analysis-resynthesis techniques, one can manipulate the parameters and study their perceptual importance. Next, the parameter settings that are most successful in conveying the intended emotion, provide optimal values that can be used in rule-based speech processing.

A very general problem with this kind of investigation, regardless of whether one focuses on perceptual aspects through listening tests or on phonatory/articulatory aspects through acoustic analysis, is the question of ecological validity. How does one know that speakers, when asked to express a certain emotion, do not simply reproduce something they acquired in training? How does one know that participants in listening experiments do not simply learn to assign a certain label to a perceived utterance, without their answer having anything to do with the real perception of a sign of affective arousal in the speech? The true answer is that we never know for sure. We can, however, take measures in the design of experiments, that minimize the risk of artificial behavior. The combined approach of eliciting affective states in speakers, and independently testing the identification

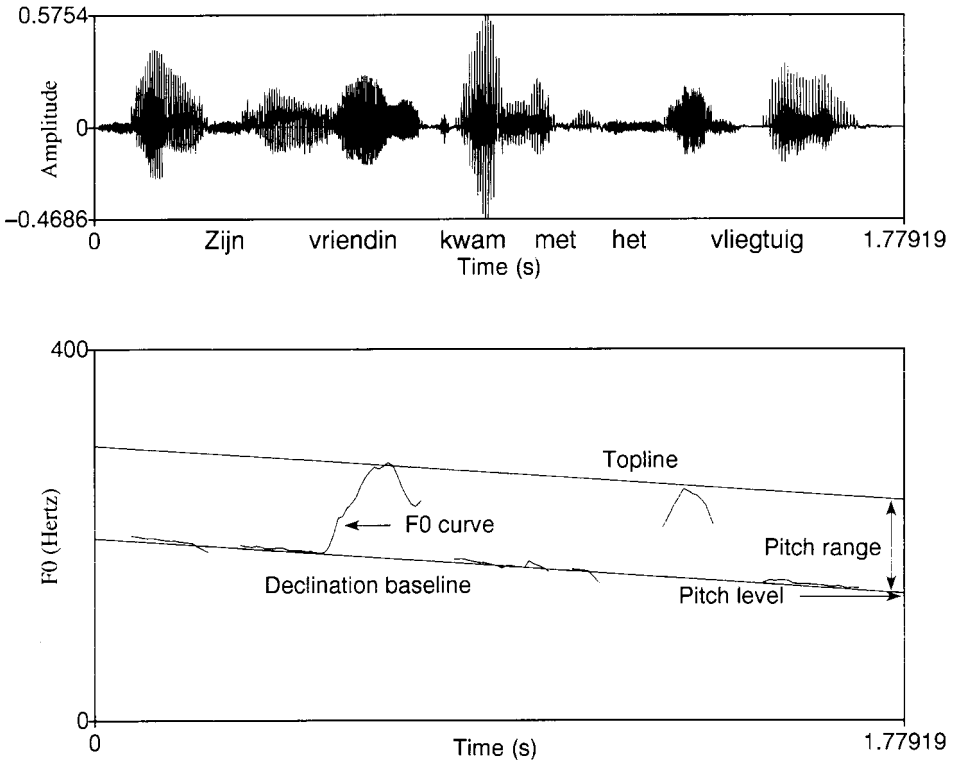

Figure 1. Representation of notions associated with pitch.
The speech sample used as example corresponds to Sentence 2: 'plane' (His girlfriend came by plane).

of these emotions by listeners, is such a measure. Furthermore, in the experiments presented here, subjects in listening tests were never given feedback about correctness or consistency of their responses. In identification experiments, there were never more than four repetitions of the same emotion realized in the same utterance, and stimuli were presented in different, random orders to each listener. Subjects took listening-test runs independently of one another, while for each experiment, a new group of listeners was engaged. Without giving an absolute guarantee, all these measures should increase the likelihood that the conclusions drawn from the experimental data apply to the same perceptual and performance behavior as expression and perception of emotions in every-day life. Furthermore, the matter of selecting adequate speech material for such a study is not a trivial one. Intuitively, the best option would be to use spontaneous emotions uttered in daily-life situations. However, the difficulty of obtaining good recordings of such material, appropriate for analysis and manipulations, labeled in terms of what was really expressed by the speaker, comparable from one emotion to the other, while respecting ethics is known to be a difficult issue. Some concessions have to be made. Considering the fact that, in this study, the intention is merely to determine a model of the listener,

and not a model of the speaker, the primary matter is that the speech material is adequate for the perception tasks. Moreover, the need for control of the sources of variability in the speech material is important in such a study about variability in speech, especially because our understanding of speech variability is still rather limited. In this light, it seems reasonable to use well-controlled speech material, appropriate for generating controlled stimuli to be used in perception tasks, and thereby limiting the sources of variability. Consequently, it is accepted that the direct observations of this speech material should not automatically be generalized to the production of spontaneous emotional speech. However, would the speech material not be representative of natural emotions, it is to be expected that the listeners would not identify the intended emotion category successfully. Therefore, in order to obtain a restricted number of utterances for this study, a selection has been made on the basis of the perceptual identification of emotion categories. Indeed, the quality of the model of the listener will not only depend on the experimental procedures, but also on the adequacy of the speech material.

1.1. Selection of acoustic parameters

Even though earlier research (e.g., Fairbanks & Pronovost, 1939; Williams & Stevens, 1972; Bouwhuis, 1974; Ladd et al., 1985; Murray, 1989; Carlson et al., 1992; Leinonen et al., 1997; Protopapas & Lieberman, 1997) has ascertained that parameters such as voice quality, loudness, rhythm, precision of articulation, and pause structure are relevant for the vocal expression of emotion, it would be convenient if, for the time being, the emotions could be modeled by a limited number of parameters. Moreover, as various interactions occur between speech parameters, it seems sensible to initiate the study of emotion in speech by considering the conveyors of major importance, conveyors that should fulfill the condition that they are used with consistency in the expression of emotion in speech. This should provide a stable basis that could be extended later by studying more detailed information on such parameters. Various studies have already investigated the role of acoustic parameters for conveying information on the affective state of the speaker. Kitahara and Tohkura (1992) asserted that pitch structure, temporal structure, and amplitude structure contribute, more than spectral structure, to the expression of emotions in speech. Frick (1985) reported that loudness was not an important means for communicating emotion. Note that the perception of loudness is affected by variations in the amplitude of the speech signal. However, the vocal effort accompanying the realization of loud speech affects more parameters than amplitude alone. Lieberman and Michaels (1962) found that F_0 is very important, but that if it is the only information, it is insufficient to transmit full emotional content. House (1990) reported that F_0 is an important cue to the perception of mood, but that other cues such as the interplay between F_0, intensity dynamics, spectral characteristics, and voice quality are also crucial to the expression of emotion.

Williams and Stevens (1972) determined that the aspect of the speech signal providing the clearest indication of the emotional state of the speaker is the 'contour of F_0 vs. time', i.e., the F_0 curve; emotion appeared to modify the pitch-curve shape generated for a breath group in several ways. Cosmides (1983) found that mean F_0 is one of the most consistently used parameters in the expression of emotion in speech. Cahn (1990) wrote that ''the acoustic features affected by emotion are mostly the F_0 and duration correlates'' (p. 38). Other studies have shown that average pitch level and average pitch range differ from one emotion to another (e.g., Fairbanks & Pronovost, 1939; van Bezooijen, 1984; Carlson et al., 1992). Moreover, Ladd et al. (1985) showed that pitch curves and voice quality have independent effects for conveying emotion in speech. This suggests that the study of pitch curves and that of voice quality can be conducted independently.

As intonation and speech rate have already been assessed to be highly relevant parameters for conveying emotion in speech, the focus will be narrowed, in the present perception study, to three parameters: pitch level, pitch range, and speech rate. These parameters are easy to control and to manipulate, also in a text-to-speech system, so that their communicative relevance can be checked. These parameters are considered at the global level of the utterance as a whole. The term 'pitch level' simply refers to how high or low the pitch, pitch being the perceptual correlate of the fundamental frequency (F_0) in the speech signal. 'Pitch range' refers to the magnitude of the F_0 fluctuations in the course of utterances. A few notions related to pitch are schematically represented in Figure 1. Considering the concrete instantiation of these parameters, the experiments were set-up within the IPO model of intonation ('t Hart et al., 1990). The IPO model is a two-component model of intonation. In this model, an F_0 curve is described as a declination baseline on which the perceptually relevant pitch movements are superimposed. Equivalently, an F_0 curve can be described as a lower and an upper declination line between which the relevant pitch movements are realized. It is thus natural, in the IPO model, to use the F_0 value at the end of the declination baseline as a measure of pitch level, and to use the excursion size of the pitch movements, i.e., the distance between declination baseline and topline as a measure of pitch range. However, in F_0 curves of natural utterances, it is not generally possible to unambiguously determine the baseline, its end frequency, and the excursion size of the pitch movements. These quantities have to be estimated on the basis of measurable parameters, such as mean pitch for estimating pitch level. As for estimating the pitch range, the best measure would be to consider the pitch variation between topline and baseline. Because of the difficulty of actually determining this declination lines in natural speech, it was decided instead, in the analysis of natural speech, to use a measure that considers the pitch variation around the mean pitch, i.e., the standard deviation of the mean pitch. Summarizing, pitch level can be instantiated by measures of mean F_0, or end frequency of the utterance, while pitch range can be instantiated by measures of the standard deviation of mean F_0 or the distance between 'F_0 minima' and 'F_0

maxima'. Finally, speech rate will be represented by the duration of an utterance relative to its duration when produced in the expression of neutrality.

1.2. Aim of the study

A first goal is to verify whether emotional synthetic utterances can be generated by rule, by means of a modification of the usual settings for rule-based synthesis. A second goal is to determine whether the parameters pitch level, pitch range, and speech rate, at the rather global level of the full utterance, are suitable for generating emotion in speech by overlaying modifications on neutral utterances. While assessing the perceptual relevance of these acoustic parameters for conveying emotion in speech, a third goal is to determine which parameter values are optimal for conveying various emotions, thereby constituting a model of the listener making only use of acoustic variations for identifying the emotion categories. The choice of generating emotions vocally by converting neutral utterances into emotional ones is based on the fact that, traditionally, most text-to-speech systems are designed to produce non-involved, neutral speech. Therefore, the focus of this study is on deviations from these basic 'neutral' settings.

In the present study, basic rules about pitch and duration are formulated, quantifying the acoustic parameters studied. Rule-driven duration, and F_0 curves with rule-driven values for pitch level and pitch range are imposed on neutral natural speech and on synthetic speech. The emotions conveyed by such utterances are tested in perception experiments. In this way, speech synthesis does not only constitute a goal in itself, but also a research tool for this investigation (Carlson, 1991; Carlson et al., 1992; Beckman, 1997).

2. Speech Material

Speech material was selected from a database recorded at IPO. This database contains speech from three Dutch speakers, two male and one female informants, who are native speakers of Dutch. They were instructed to produce thirteen emotions: neutrality as a category of reference, joy, happiness, boredom, worry, anger, sadness, fear, guilt, disgust, haughtiness, indignation, and rage. These emotions were elicited in evocative situations, i.e., the speakers used sentences of semantically emotional content, such as, 'How nice to see you here' for the expression of joy, and once they felt in the intended mood, they spoke out a fixed group of eight sentences. This was done for each emotion, and the procedure was repeated three times per speaker. The 936 speech samples which thus formed the database (3 speakers $\times$ 8 sentences $\times$ 13 emotions $\times$ 3 times) were digitized with 16-bit precision, at a sampling frequency of 10 kHz. The eight sentences were:

(1) 'Zij hebben een nieuwe auto gekocht' (They have bought a new car),
(2) 'Zijn vriendin kwam met het vliegtuig' (His girlfriend came by plane),

(3) 'Hij is morgen naar Amsterdam' (He is in Amsterdam tomorrow),
(4) 'Het is bijna negen uur' (It is almost nine o'clock),
(5) 'Zij heeft gisteren gebeld' (She phoned yesterday),
(6) 'De lamp staat op het bureau' (The lamp is on the desk),
(7) 'Jan is naar de kapper geweest' (John has been to the hairdressers),
(8) 'Zij was aan het telefoneren' (She was making a phone call).

The content of these sentences was intended to be semantically neutral with respect to emotion.

A preliminary perception experiment has led to a selection of the thus obtained speech material on the basis of identification performances. This resulted in fourteen utterances, two sentences spoken with seven emotions by one of the male speakers. The male speaker and the two sentences were selected because these two sentences spoken by him yielded the best identification. The seven emotions were selected because they were identified correctly in at least 50% of all trials. The two selected sentences were Sentence 1: 'car': 'Zij hebben een nieuwe auto gekocht' (They have bought a new car), and Sentence 2: 'plane': 'Zijn vriendin kwam met het vliegtuig' (His girlfriend came by plane). The seven selected emotions were neutrality (Dutch category name: 'neutraliteit'), joy ('blijheid'), boredom ('verveling'), anger ('boosheid'), sadness ('verdriet'), fear ('angst'), and indignation ('verontwaardiging').

3. Frame of Reference

3.1. Experiment 1: Identification of the Emotions in the Natural Utterances

The first experiment was carried out in order to provide a baseline for the identification of the emotions in the fourteen selected utterances. Its purpose was to determine how well the emotion could be identified in each selected natural utterance, and how good an example of a particular emotion the subjects found each utterance to be.

3.1.1. *Procedure*

The 14 selected utterances (1 speaker $\times$ 2 sentences $\times$ 7 emotions) serving as stimuli were presented to the subjects in two blocks, one block per sentence. Each block contained 14 trials, as each speech sample was presented twice. Stimuli were presented to each subject in a different random order. Sentence order was counterbalanced across subjects. This design resulted in 28 stimuli. As motivated in the introduction, there was no training or feedback, as is also the case in all the following experiments reported in this study. After hearing an utterance over headphones, the subjects had to choose among the labels of the seven emotions, the one that best corresponded to the emotion conveyed by the utterance. After this forced choice from the seven alternatives, subjects also attributed an adequacy rating for that emotion, ranging between 1 (bad) and 5 (good). These adequacy ratings indicate

how well the utterance conveys the emotion. Ten subjects participated in the experiment.

3.1.2. *Results*

When the attributed emotion label corresponded with the emotion the speaker intended to communicate, the response was considered to be correct. For each sentence, the proportion of correct responses was then determined. Furthermore, a mean adequacy rating was computed by averaging the adequacy ratings over the correct responses. Mean proportions of correct responses and mean adequacy ratings are presented in Table I, for both sentences separately. The confusion matrix of the results pooled over both sentences is given in Table II.

Fear and anger were less successfully identified than the other emotions. A positive statistically significant correlation was found between the adequacy ratings and the proportion of correct responses [$r(14) = 0.69$, $p < 0.007$]. This indicates that utterances in which the emotion was easy to identify, were also considered to be more appropriate for that emotion.

Table I. Mean proportion of correct responses and mean adequacy ratings on a 5-point scale for Experiment 1: identification of the emotions in the natural utterances selected

Emotion	Sentence 1: 'car' (They have bought a new car)	Sentence 2: 'plane' (His girlfriend came by plane)	Average over both sentences
Neutrality	1.00 (4.15)	0.90 (4.06)	0.95 (4.11)
Joy	0.90 (3.11)	0.55 (2.91)	0.73 (3.02)
Boredom	0.95 (3.95)	0.95 (3.95)	0.95 (3.95)
Anger	0.50 (3.20)	0.35 (2.71)	0.43 (2.96)
Sadness	0.95 (4.05)	0.90 (4.39)	0.93 (4.22)
Fear	0.30 (3.83)	0.35 (3.14)	0.33 (3.48)
Indignation	0.80 (4.06)	0.95 (4.21)	0.88 (4.14)
Mean	0.77 (3.76)	0.71 (3.62)	0.74 (3.70)

Table II. Confusion matrix for Experiment 1: identification of the emotions in the natural utterances

Intended Emotion	Responses of subjects							
	Neutrality	Joy	Boredom	Anger	Sadness	Fear	Indignation	Total
Neutrality	**38**	0	2	0	0	0	0	40
Joy	7	**29**	0	2	0	0	2	40
Boredom	0	0	**38**	0	1	0	1	40
Anger	9	7	0	**17**	0	1	6	40
Sadness	0	0	0	1	**37**	2	0	40
Fear	2	8	0	1	2	**13**	14	40
Indignation	0	5	0	0	0	0	**35**	40

Results are pooled across the two presentations of the two sentences and the ten subjects.

3.1.3. *Discussion*

The results form a frame of reference against which the results of the following experiments will be compared. Some emotions appeared to be more difficult to identify than others. The differences in adequacy ratings suggest that the speaker was not always equally successful in conveying the different emotions. According to the adequacy ratings, the least successfully conveyed emotions were found to be anger and joy. Indeed, it is possible that some emotions are intrinsically more difficult to identify from the speech signal only than others. This seems quite reasonable considering that the physiological reactions caused by different emotions might be quite similar, which, in the acoustic space, could lead to a substantial overlap of the sets of acoustic correlates of different emotions. Emotions such as fear and anger can, for example, both provoke dryness of the mouth (Williams & Stevens, 1972). Also, a specific emotion might provoke various physiological states, which could lead to varying acoustic correlates. Different physiological states might also lead to acoustic correlates lying close to each other in acoustic space. Fear appeared to be confused quite frequently with indignation, while no utterances intending to convey indignation were confused with fear. A possible explanation is that the set of speech correlates for indignation could be more scattered than the set for fear, with an overlap between the two sets. Neutrality and boredom, on the other hand, are hardly ever mistaken for any other emotion. Considering the fact that neutrality has been chosen as a reference to which all selected categories relate, this observation concerning neutrality is reassuring. Confusion of other emotions with boredom is also very rare. This suggests that the acoustic correlates for boredom are distinctive. Furthermore, the positive correlation between identification rates and 'adequacy values' suggest that the identification rate is not merely a measure of discrimination among stimuli, but also reflects the successfulness of conveying the intended emotion.

3.2. Experiment 2: semantic content

Although aware of the fact that there might always be a certain interaction between the linguistic content of a sentence and the conveyed emotion, an attempt was made to put an emotive overlay on a sentence which is semantically as neutral as possible. The semantic content of the utterances might still bias the responses of the subjects, who could experience greater facility or difficulty in associating specific emotions with particular sentences. The purpose of this control experiment was to test whether the semantic content of the utterances had any impact on the identification of the emotions.

3.2.1. *Procedure*

The two written sentences used in the present study formed the experimental material. Twenty subjects participated in the experiment. The two sentences were presented on paper to the subjects who were asked to read silently and to indicate,

on a scale from 1 (bad) to 5 (good), how well the semantic content of each sentence fitted each of the seven given emotion categories. The mean 'semantic adequacy rating' and its standard deviation were computed.

3.2.2. *Results*

Ideally, suitable sentences would not only be of semantically neutral content, but would also allow the expression of each of the emotions included in the study. Therefore, the expected results consist of a high score for neutrality, and equally low, but not zero, scores for all other emotions. The mean semantic adequacy ratings and its standard deviation are reported in Table III. It appears, indeed, that for both sentences, neutrality received the highest score on the semantic adequacy scale. On the other hand, the ratings obtained for the other emotions are not all equally low, which indicates that the emotions are not all equally likely to be conveyed by merely the semantic content of the sentences. The ratings obtained for joy, for instance, while remaining lower than for neutrality, are higher than for the other emotions. However, the most important point is whether the semantic content of the sentences influences the identification of the emotions. In order to test this point, the semantic adequacy ratings were compared with the results of Experiment 1. If the semantic content of the sentences is partially responsible for the differences in identification of the emotions in Experiment 1, one would expect the identification rates in Experiment 1 to correlate positively with the semantic adequacy ratings. On the other hand, if semantic adequacy and identification rates are independent from each other, correlation would be zero. It appears that the semantic adequacy correlated neither with the identification rates of Experiment 1 [$r(14) = 0.33$, NS], nor with the adequacy ratings of Experiment 1 [$r(14) = 0.00$, NS]. The two correlation coefficients do not significantly deviate from 0. This lack of correlation with the results from Experiment 1 shows that, despite the variations in the scores, the semantic content of the sentences had no substantial impact on the identification scores of the emotions in the perception experiment. Therefore, these sentences are considered suitable for use in the present study.

Table III. Mean semantic adequacy ratings on a 5-point scale and standard deviation (in parentheses) for Experiment 2: semantic content

Emotion	Sentence 1: 'car' (They have bought a new car)	Sentence 2: 'plane' (His girlfriend came by plane)
Neutrality	3.90 (1.41)	4.25 (1.12)
Joy	3.30 (1.42)	2.80 (1.11)
Boredom	1.50 (0.83)	1.45 (0.83)
Anger	1.85 (0.99)	1.45 (0.76)
Sadness	1.25 (0.44)	1.40 (0.82)
Fear	1.15 (0.49)	1.80 (1.15)
Indignation	2.35 (1.35)	1.95 (1.05)

Means are pooled across the two sentences and twenty subjects.

3.3. Prosodic analysis of the speech material

In order to obtain a first approximation of the selected acoustic parameters at utterance level: overall speech rate, pitch range, and pitch level, an analysis was carried out on the 14 utterances selected. The speech rate was represented by the overall utterance duration relative to neutrality. The pitch curve was determined by subharmonic summation (Hermes, 1988). As mentioned in the introduction, the pitch level was represented by the overall mean F_0, and the pitch range by the standard deviation of this mean. Later, we will calculate how well these measures correspond with the measures used for pitch level and pitch range within the IPO two-component model of intonation, i.e., the end of the declination baseline and the distance between the declination baseline and topline, respectively.

Tables IV and V present the absolute duration, the relative duration with respect to the neutral utterance of the corresponding sentence, the mean pitch values, and the standard deviation of the pitch values. The expression of boredom and, to a lesser extent, indignation is realized with a low speech rate, while fear and anger were expressed with rather high speech rates. Note that a high speech rate results in rather short utterances, and a low speech rate in rather long utterances. Neutrality and boredom were conveyed with rather low average pitch values; while rather high

Table IV. Duration, pitch, and intonation patterns of Sentence 1: 'car' (They have bought a new car)

Emotion	Duration (sec)	Duration relative to neutrality	Mean pitch in Hz	S.d. pitch	Intonation pattern
Neutrality	1.66	1.00	132	18.7	15&A
Joy	1.58	0.95	205	30.6	1D3C
Boredom	2.58	1.55	131	11.0	1D3&A
anger	1.45	0.87	201	31.6	15&A
Sadness	1.84	1.10	168	12.7	14E
Fear	1.40	0.84	237	19.2	3C
Indignation	1.96	1.18	254	37.8	45&A

Table V. Duration, pitch, and intonation patterns of Sentence 2: 'plane' (His girl friend came by plane)

Emotion	Duration (sec)	Duration relative to neutrality	Mean pitch in Hz	S.d. pitch	Intonation pattern
Neutrality	1.76	1.00	135	19.2	1D1D
Joy	1.78	1.01	205	39.3	1D3C
Boredom	2.35	1.33	133	13.2	1D1&A
Anger	1.58	0.89	193	45.3	15&A
Sadness	2.12	1.20	170	19.2	1D1D
Fear	1.53	0.86	230	27.2	1D3C
Indignation	2.35	1.33	245	53.3	45&A

values were produced for fear and indignation. For indignation, anger, and, to a lesser extent, joy, the rather large standard deviation of the mean pitch seems an indication of rather large pitch variations.

In addition, in order to describe the shape of the pitch curves realized in each utterance, a perceptual analysis of the pitch curves was carried out. This consisted of labeling the intonation of each utterance according to the Dutch intonation grammar by 't Hart et al. (1990). Basic units of this grammar are pitch movements, such as 'early prominence-lending rise' ('1') or 'late prominence-lending fall' ('A'). In this intonation transcription, digits refer to rises and letters to falls. A sequence of pitch movements for the whole utterance, if at least legal according to the grammar of intonation, constitutes an intonation pattern. Each utterance was thus attributed an intonation pattern. The results of this labeling are presented in Tables IV and V. Although, in specific emotions, the same intonation pattern was realized for both sentences, no unique relationship was found between specific emotions and specific intonation patterns.

4. Optimal Values for Pitch Level, Pitch Range, and Speech Rate

Now that acoustic parameters have been selected and a reference has been established, the following series of experiments will determine a model of the listener, i.e., optimal perception values for pitch level, pitch range, and speech rate, to be used for rule-based generation of emotional speech. Although it is not tested in the present study, it is to be expected that the same values can also be used for rule-based identification of emotion. The term 'optimal' is used for values that, among all values proposed in the experimental set-up, provide the best results in identification tests. In order to avoid caricatural expression of emotions, the experimental set-up did not include parameter values resulting in an unnatural expression of emotion. In Experiment 3, an attempt will be made to determine optimal values for pitch level and pitch range that will be tested against the natural pitch curves in Experiment 4. Experiment 5 will aim to determine optimal values for speech rate. In Experiment 6, the optimal values will be tested in re-synthesized speech. Finally, Experiment 7 will be an ultimate test for the optimal values in synthetic speech. These values will make up rules for transforming neutral speech into emotional speech.

4.1. Experiment 3: optimal values for pitch levels and pitch range

The present experiment aims at finding optimal perception-based parameter values for pitch level and pitch range, for each emotion. While seeking these optimal values, variation in shape of the pitch curves, i.e., intonation patterns, will be controlled for.

4.1.1. *Procedure*

Per natural utterance, nine synthetic F_0 curves, systematically varied in pitch level and in range, were generated[1]. These synthetic curves, all synthesized with the same intonation pattern and segmental duration, were transplanted on the corresponding utterances. For each sentence, the 9 variants per emotion (3 end frequencies $\times$ 3 excursion sizes) served as stimuli.

The two sentences were presented in blocks. Sentence order was counterbalanced across subjects. In a block, the stimulus order was randomly varied per listener. Ten subjects, different from the ones in the previous experiments, participated in the listening test. For each given emotion label, subjects could listen to the nine 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 selection one point. Per sentence, the mean score for each intonation variant was calculated.

4.1.2. *Results*

The parameter values that led to the highest mean score were considered to be optimal for the specific emotion. These values and the corresponding mean score are presented, per emotion, in Table VI. When both sentences are compared, a good agreement in the optimal values is shown. Spearman's rank-correlation test yields a high correspondence between the pitch levels (rs = 0.9821) and between the pitch ranges (rs = 0.9196) in both sentences. The results suggest that the parameters representing pitch level and pitch range, that are usually attributed standard values in speech synthesizers, are perceptually relevant for the generation of vocal emotions.

[1]Rule-based F_0 curves, which replaced the original F_0 curves of the emotional utterances, were generated on the basis of general rules for the synthesis of Dutch intonation. This was done using dedicated software (Zelle et al., 1984) for producing nine synthetic F_0 curves per natural utterance. Two parameters were systematically varied: end frequency of the baseline in Hertz (Hz), which determines the pitch level, and excursion size of the pitch movements in semitones (s.t.), which determines the pitch range. The synthetic F_0 curves varied over three excursion sizes and three end-frequency values. The variation in excursion size involved steps of 2 semitones around the values produced by the speaker in the natural utterances, and the variation in end frequency involved steps of 15 Hz around the values observed in the natural utterances. Note that, in standard applications for neutral speech synthesized in Dutch, the end frequency is fixed at 75 Hz, and the size of the pitch movements is fixed at six semitones (Collier, 1991). Per emotion, the intonation patterns produced in Sentence 1 'car' (see Table IV) were used for synthesizing the pitch contours for both sentences. In this way there was no difference in intonation pattern between speech samples of the same emotion. Naturally, the exact location of the vowel onset of the accented syllable was determined individually for each speech sample. The transplantation of the synthetic F_0 curves onto the natural emotional utterances was based on the Time-Domain Pitch-Synchronous OverLap-and-Add (TD-PSOLA) algorithm (Charpentier & Moulines, 1989; Verhelst & Borger, 1991). In this experiment, the manipulation copied the segmental duration of the natural utterances.

Table VI. Optimal parameter values for the rule-based pitch curves with corresponding scores and their standard deviations

		Sentence 1: 'car' (They have bought a new car)			Sentence 2: 'plane' (His girlfriend came by plane)		
Emotion	Intonation pattern	Pitch level (End frequency)	Pitch range in semitones	Score (s.d.)	Pitch level (End frequency)	Pitch range in semitones	Score (s.d.)
Neutrality	15&A	65	6	1.90 (1.20)	65	4	2.10 (0.99)
Joy	1D3C	155	10	2.40 (1.26)	155	10	2.90 (0.32)
Boredom	1D3&A	65	4	3.00 (0.00)	65	4	2.20 (1.14)
Anger	15&A	110	10	2.70 (0.95)	110	10	2.30 (1.25)
Sadness	14E	95	6	1.30 (1.25)	110	8	1.60 (1.51)
Fear	3C	200	8	2.30 (0.95)	200	7	1.80 (1.03)
Indignation	45&A	155	10	2.40 (1.26)	185	10	2.50 (1.08)

Note that a score of 1 means on average a third place out of nine variants, while a score of 3, i.e., the maximum score, means the best choice out of the nine variants.

The intonation pattern may also constitute relevant information. The utterances found to convey best the expression of joy and indignation in Sentence 1, were synthesized with the same optimal pitch level and pitch range values. However, these emotions were produced with structurally distinct intonation patterns. On the other hand, the same intonation pattern was used in utterances conveying best different emotions, i.e., neutrality and anger in Sentence 1. The corresponding utterances were generated with different pitch level and pitch range values. These considerations suggest that the intonation pattern also plays a role in conveying emotion vocally (Mozziconacci & Hermes, 1999), which justifies the control of this element in the present study.

It also has to be noted that, in many cases, the versions that received the highest scores have extreme values (highest versus lowest frequency, smallest versus largest excursion). The fact that extreme values were selected by the subjects raises the question whether for some emotions, subjects would have chosen even more extreme values if those had been included in the set of stimuli. On the other hand, emotions generated using more extreme values could have been perceived as exaggerated. Such cases have been avoided on purpose, by limiting, in the manipulations, the extent of the deviations from the values produced in the natural speech (steps of 15 Hz for pitch level, and 2 semitones for pitch range). Furthermore, it can be observed that subjects agreed more on the optimal realization of some emotions than on the optimal realization of some others. This is represented by the standard deviation of the adequacy ratings (Table VI).

4.2. EXPERIMENT 4: TEST OF THE OPTIMAL PITCH CURVES

The present experiment tests the adequacy of the values found optimal in Experiment 3 for generating emotional speech, and investigates the effect of manipulating the

pitch curve. Indeed, as rule-based synthetic pitch curves were generated according to the principles of the Dutch intonation grammar by 't Hart et al. (1990), they are, in fact, made of straight lines in the log F_0 vs. time domain. In order to be sure that this way of generating the pitch curve is not a factor, the rule-based contours, generated with the optimal values for pitch level and pitch range obtained from Experiment 3, are compared with stylized versions of the original pitch curves in the natural emotional utterances.

4.2.1. *Procedure*

Test utterances were made by transplanting the synthetic pitch curves[2] found to be optimal in Experiment 3, onto the natural emotional utterances. Ten new subjects participated in the experiment. For both sentences, two blocks of fourteen utterances were successively presented to the subjects. Each block contained seven close-copies obtained by stylization of the original utterances, and seven utterances with a rule-based F_0 curve. Stimuli were presented to each subject in a different random order. The sentence order was counterbalanced across the subjects. After listening to an utterance once, subjects had to choose one of the seven labels corresponding to the emotion they thought was conveyed in the utterance. In contrast with previous experiments, subjects were not asked to supply adequacy ratings, because utterances with stylized pitch curves were also involved. These utterances were perceptually indistinguishable from the original, thereby inducing a strong adequacy bias. As a consequence, the adequacy ratings for the manipulated versions were postponed until a later experiment.

4.2.2. *Results*

The proportion of correct responses was computed per subject. A two-way analysis of variance (ANOVA), with the seven emotions and the two types of pitch curves as within-subjects variables, showed that the effect of type of pitch curve was not significant [$F(1,9) = 2.17$, $p = 0.175$]. The emotions in the utterances with the rule-based pitch curve were identified almost as well as the ones with the close-copy stylization of the original utterances. The effect of emotion was very significant [$F(6,54) = 9.76$, $p < 0.001$]. An interaction was found between pitch, i.e., pitch level and pitch range, and emotion [$F(6,54) = 2.58$, $p = 0.029$]. The mean proportion, reported in Table VII, was pooled over the two sentences. The confusion

[2]The rule-based synthetic pitch curves are a stylization of the F_0 curve, made of straight lines in the log F_0 vs. time domain, and realized with the values for pitch range and pitch level that received the highest scores in the previous experiment. In order to be sure that stylizing is not a factor, control utterances also had stylized pitch curves, but these were merely close-copies (de Pijper, 1983) of the natural utterances. Indeed, a close-copy stylization is a straight-line approximation of the F_0 curve, which, in re-synthesis, is perceptually indistinguishable from the original, and has the minimum number of straight-line segments. This second set of utterances was re-synthesized without any manipulation of pitch level or pitch range. All the manipulations were again done by means of the PSOLA techniques. The temporal structure and the voice quality of the natural emotional utterances were left unchanged.

Table VII. Mean proportion of correct responses for the original and the rule-based pitch curves

	Responses of subjects							
Pitch curves	Neutrality	Joy	Boredom	Anger	Sadness	Fear	Indignation	Mean
Original	0.85	0.62	0.92	0.32	0.97	0.60	0.85	0.73
Rule-based	0.67	0.72	0.85	0.42	0.75	0.42	0.77	0.66

Table VIII. Confusion matrix for Experiment 4: test of the optimal pitch curves

Intended	Responses of subjects							
Emotion	Neutrality	Joy	Boredom	Anger	Sadness	Fear	Indignation	Total
Neutrality	**61**	0	10	3	3	0	3	80
Joy	10	**54**	1	3	1	5	6	80
Boredom	3	0	**71**	6	0	0	0	80
Anger	31	15	0	**30**	0	0	4	80
Sadness	5	0	3	2	**69**	0	1	80
Fear	2	7	0	0	10	**41**	20	80
Indignation	0	1	0	6	5	3	**65**	80

Results are pooled across two presentations of the two types of pitch curves, on the two sentences for the ten subjects.

matrix is presented in Table VIII. All emotions were identified well above chance level (0.14), though anger (0.42) and fear (0.42) were less well recognized than the other emotions. As can be seen in Table VIII, utterances expressing fear were often labeled as indignation, and anger was very frequently confused with neutrality.

4.2.3. *Discussion*

The results in Table VII show that the difference in identification of the emotions in the test utterances and the control utterances was rather small; the Spearman's rank-correlation [rs = 0.8036] indicates a high correspondence between the results obtained with both types of stimuli. This suggests that the manipulation had a small, though significant ($t_s = 1.976$, df = 158, $p < 0.05$), deteriorating effect. The results show that the complex original F_0 curve of an emotional utterance can be replaced by a simple approximation which only requires a limited number of parameters: excursion size which determines the pitch range, end frequency which determines the pitch level, and intonation pattern which determines the shape of the pitch contour. Moreover, the intonation patterns seem to be suitable for the generation of emotional speech. All emotions were identified well above chance level, in fact, fairly well, but it should be remembered that the utterances had the duration and information on the voice quality from the original utterances, so that the fairly good identification of the emotions is partly due to the presence of these other features of speech.

4.3. EXPERIMENT 5: OPTIMAL SPEECH RATE

This experiment is concerned with overall speech rate. Temporal variations can be realized as variations in articulation rate, rhythm, pause structure, or other temporal features. Clearly, the simplest manipulation of speech rate consists of a linear time compression or expansion of whole utterances, modifying the overall duration of the utterances, including both speech and any pauses within utterances, without affecting the fine temporal structure of the speech. If this linear approach, in which speech rate is inversely proportional to the overall utterance duration, is not too global for the expression of emotion in speech, optimal speech rates relative to neutrality may be determined for each emotion. These optimal values for speech rate could then also be applied in rule-based diphone speech.

4.3.1. *Procedure*

Stimuli were generated by means of speech manipulations[3]. First, per emotion, for both previously used sentences, the neutral utterance was provided with the F_0 curve and the overall duration of the emotional speech samples. Second, seven temporal variants were created by compressing/expanding the overall resulting utterances by 70, 80, 90, 100, 110, 120, and 130 percent.

Ten subjects, different from the previous ones, participated in the experiment. The versions resulting from the overall speech rate manipulations were organized into two blocks, i.e., one for each text, and randomized. The sentence order was counterbalanced across listeners. The procedure was exactly the same as Experiment 3, except that the subjects now had to choose and rank the three versions they found to be the best among the seven duration versions (instead of nine pitch versions in Experiment 3).

4.3.2. *Results*

As in Experiment 3, on the basis of the rank-ordering, a score of three points was assigned to the best variant, two points were assigned to the second best variant and one point to the third best. The mean score of each variant was computed. The variant that received the highest mean score was considered to be optimal

[3]The two original neutral utterances served each as carrier utterance for the corresponding sentence. First, they were provided with the F_0 curve of the emotional speech samples, time aligned by means of the previously mentioned TD-PSOLA algorithm (Verhelst & Borger, 1991). In order to copy this time-aligned F_0 curve, the optimal Dynamic Time Warping (DTW) path was calculated between the emotional and neutral utterance, 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 precision of this time-domain manipulation is limited to an integer number of pitch periods. The resulting utterances, thus, had the same pitch curve as the expressive 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. These seven variants per emotion served as stimuli.

Table IX. Optimal sentence duration (speech rates) relative to neutrality and scores

	Sentence 1: 'car' (They have bought a new car)		Sentence 2: 'plane' (His girlfriend came by plane)		
Expression	Relative sentence duration (speech rate)	Score (s.d)	Relative sentence duration (speech rate)	Score (s.d.)	Mean relative sentence duration (mean speech rate)
Neutrality	100% (1.00)	2.00 (1.00)	100% (1.00)	2.40 (0.92)	100% (1.00)
Joy	85% (1.18)	2.30 (0.64)	80% (1.25)	2.00 (1.18)	83% (1.20)
Boredom	154% (0.65)	1.70 (1.42)	145% (0.69)	1.90 (0.94)	150% (0.67)
Anger	78% (1.28)	1.40 (0.92)	80% (1.25)	1.90 (1.14)	79% (1.27)
Sadness	126% (0.79)	1.80 (0.92)	131% (0.76)	2.20 (0.87)	129% (0.78)
Fear	92% (1.09)	1.50 (1.02)	85% (1.18)	1.50 (1.02)	89% (1.12)
Indignation	129% (0.78)	1.80 (0.87)	106% (0.94)	1.40 (1.20)	117% (0.85)

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.

for that particular emotion. The corresponding optimal relative sentence duration, with score and standard deviation are reported in Table IX. This sentence duration relative to neutrality was calculated for each emotion, by dividing the mean overall duration for a specific emotion by the sentence mean for neutrality. 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.

4.3.3. *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. For indignation, the optimal overall duration varied from 106% to 129% (see Table IX). For this emotion, the subjects showed a tendency to select rates around the mean value. These results suggest that a variation in speech rate is easily tolerated for the expression of indignation, possibly because speech rate is not a very important attribute for conveying indignation. In addition, the results of the present experiment correspond rather well with the speech rates produced in the original samples (see Tables IV and V) which corroborates the adequacy of the values found here.

The logical next step would be to test the speech rate values just obtained. It did not seem very interesting, however, to test speech rate values prior to testing both intonation and speech rate optimal values together. Indeed, letting subjects attribute an emotion label to neutral utterances on which only a manipulation of speech rate is imposed, did not seem sensible; under neutral circumstances, time-pressure would probably be considered to be the most important factor determining speech rate, and time-pressure is not what was investigated here. Another possibility was to test the optimal speech rate values by imposing them on emotional utterances, and comparing the identification performance against the one obtained with the original

speech rate values. This option did not seem useful either, especially as the optimal values were only searched around the values produced by the speaker in the natural speech. This would have resulted in rather small differences between the two conditions. Such a test would therefore only serve the purpose of confirming the role of speech rate for the expression of emotion in speech. Moreover, identification of the emotions in the natural emotional utterances was already so high that a ceiling effect could have occurred. For this reason, rules concerning pitch and duration were tested together in the following experiment. The mean optimal values for pitch level and pitch range found for each emotion in Experiment 3, and the mean optimal values for speech rate found in the present experiment, were used simultaneously in an attempt to generate emotional speech from neutral speech.

4.4. Experiment 6: testing optimal values for pitch level, pitch range, and speech rate on manipulated re-synthesized neutral speech

This experiment investigated the adequacy for conveying emotion in speech, of the values for pitch level, pitch range, and speech rate found to be optimal for each emotion in the present study. Doing so, it investigates whether it is possible, by applying these values, to generate emotional speech from a neutral utterance.

4.4.1. *Procedure*

A neutral utterance of both sentences was linearly compressed or expanded by means of PSOLA, according to the optimal values obtained from Experiment 5. Using the appropriate software (Zelle et al., 1984), the rule-based F_0 curve with the optimal pitch level and pitch range from Experiment 3, was generated for each particular emotion. The neutral utterances with the rule-based overall speech rate were then provided with these F_0 curves, also by means of PSOLA. No manipulation of voice quality or micro-duration structure were carried out, so that all the signals still carried this type of information from the original neutral sentences.

Ten new listeners participated as subjects in the experiment. For each sentence, two blocks of 14 trials were constructed. Per block, the seven emotional utterances were presented twice, in a different random order. The sentence order was counterbalanced across the subjects. The task of the subjects was to listen to each stimulus and to choose from the seven labeling alternatives, the one they thought was conveyed by the utterance. They were also instructed to give an adequacy rating on a scale from 1 (bad) to 5 (good) for the chosen emotion.

4.4.2. *Results*

The mean proportions of correct responses and the mean adequacy ratings, pooled across the two sentences and the ten subjects, are presented in Table X. On average, 48% of the emotions were correctly identified. In order to facilitate comparison, the reference results obtained in Experiment 1 are also presented in Table X. An ANOVA, with the seven emotions as within-subject variables, showed that the dif-

Table X. Mean proportion of correct responses and mean adequacy ratings for the pitch curves based on the optimal pitch level, pitch range, and speech rate, in comparison with the original utterances

Rule-based results (Exp. 6)

	Responses of subjects							
	Neutrality	Joy	Boredom	Anger	Sadness	Fear	Indignation	Mean
Proportion correct	0.80	0.38	0.90	0.38	0.20	0.23	0.50	0.48
Adequacy ratings	3.87	3.47	4.14	3.33	3.75	3.44	3.05	3.58

Reference results (Exp. 1)

	Responses of subjects							
	Neutrality	Joy	Boredom	Anger	Sadness	Fear	Indignation	Mean
Proportion correct	0.95	0.73	0.95	0.43	0.93	0.33	0.88	0.74
Adequacy ratings	4.11	3.02	3.95	2.96	4.22	3.48	4.14	3.70

Table XI. Confusion matrix for Experiment 6: testing optimal values for pitch level, pitch range, and speech rate on manipulated re-synthesized neutral speech

Intended	Responses of subjects							
Emotion	Neutrality	Joy	Boredom	Anger	Sadness	Fear	Indignation	Total
Neutrality	**32**	0	4	3	0	0	1	40
Joy	2	**15**	0	3	1	6	13	40
Boredom	2	0	**36**	1	1	0	0	40
Anger	16	4	1	**15**	0	1	3	40
Sadness	1	0	28	0	**8**	0	3	40
Fear	2	5	0	2	9	**9**	13	40
Indignation	4	3	2	0	4	7	**20**	40

Results are pooled across the two presentations of the two sentences and the ten subjects.

ferences between the emotions were significant [$F(6,54) = 9.76$, $p < 0.001$]. Sadness and fear were clearly identified less well than the other emotions. The confusion matrix presented in Table XI shows that this was due to the confusion of fear with indignation and sadness, and to a major confusion of sadness with boredom. The modeling of fear and sadness seems insufficient.

A comparison of the results of the present experiment, with the outcome of Experiment 1 that was concerned with the identification of emotion in natural speech (see Table X), shows that the identification of the emotions boredom (0.90), neutrality (0.80), and anger (0.38), is rather good in this experiment, if we compare it to the identification in natural speech (0.95, 0.95, and 0.43, respectively). Other emotions such as joy (0.38) and sadness (0.20) compare less successfully to the identification results obtained with natural emotional speech in Experiment 1 (0.73 and 0.93, respectively). Considering the overall picture, Spearman's rank-correlation indicates a significant but small difference ($p > 0.05$) between both series of results

for correct identification [rs = 0.6250], and a low correlation of the adequacy ratings in both experiments [rs = 0.2143]. This suggests that even if the identification of emotions, in emotional re-synthesized speech generated from neutral speech, is relatively successful, the adequacy ratings are attributed differently by the subjects in both experiments.

4.4.3. *Discussion*

Despite the difficulties encountered with some emotions, the results for all emotions stayed above the chance level of 14.3%. Clearly, the rules based on the optimal values found for each emotion are successful to a certain extent, but there are costs associated with this rule-based generation of emotional-sounding speech from neutral speech.

In particular, the expression of sadness, using the selected parameters, seems to raise difficulties. This category seems to have a distinctive set of acoustic correlates, as suggested by the high identification rate and the low number of confusions obtained with the natural utterances in Experiment 1 (see Table III). However, in the present experiment, the voice source and micro-duration features from the originally neutral utterances were present in the stimuli. This may indicate that the voice source information, the micro-duration features, or both, are required in order to convey sadness more clearly. In fact, simply listening to the sad natural utterances suggests that the voice quality is a characteristic element of this emotion. For fear, which was poorly identified in the rule-generated utterances as well as in the original speech samples, it is possible that the speaker did not attribute acoustic cues that are distinctive enough for conveying fear. Generally speaking, the identification in the rule-based versions dropped moderately in comparison to the identification in the original utterances, which is in fact what could be expected. The consideration of additional speech parameters may account for this difference.

4.5. Experiment 7: rule-based generation of expressions from diphone-concentrated synthetic speech

The previous experiments demonstrated that it is possible to generate emotional speech by means of pitch and speech rate manipulations imposed on a neutral utterance. The aim of the present experiment is to investigate whether the rules based on these optimal values convey emotion in speech if imposed on synthetic speech. The goal is to generalize the findings to speech other than the one of the original speaker. Therefore, the values for pitch level, pitch range, and speech rate were applied to synthetic diphone-speech. Sufficient identification of the emotions from diphone speech would confirm that the parameters used are powerful cues for the expression of emotion in speech.

4.5.1. *Procedure*

Utterances were synthesized[4] from a phonetic description of the two sentences, and were provided with the previously determined optimal speech rate (see Table IX) and with pitch contours computed with the previously determined optimal pitch range and the pitch level (see Table VI).

Twelve new subjects participated in the experiment. For each sentence, two blocks were realized, i.e., one per set of diphones. Each stimulus was presented twice in a block. The speech samples within blocks were presented in random order. The order of presentation of the sentences and the sets of diphones was counterbalanced across subjects. Before starting the test with the speech generated with a set of diphones, the subjects could get accustomed to that diphone speech by listening to a passage of 40 seconds of instructions, synthesized with the corresponding set of diphones. This experiment was once again based on a seven-alternative forced choice paradigm concerning the seven emotion labels. The speech samples were recorded on a digital audio tape for presentation to listeners.

4.5.2. *Results*

The mean proportions of correct responses, pooled across the two sentences, are presented, separately for each diphone set, in Table XII. The confusion matrix of the responses, pooled across sentences and diphone sets, is presented in Table XIII. On average, 63 percent of the emotions were correctly identified. An ANOVA, with seven emotions and two speakers as within-subject variables, yielded an almost significant scale value difference between the two diphone sets [60% vs. 67%, $F(1,11) = 2.47$, $p = 0.14$]. There were significant differences between the emotions [$F(6,54) = 11.45$, $p < 0.001$], and an interaction between the emotions and the diphone sets [$F(6,66) = 3.72$, $p < 0.005$]. Inspection of Table XII suggests that joy was better identified with diphone set 1, whereas neutrality, anger, and sadness were better identified with diphone set 2. All emotions were recognized far above the chance level of 14.3%. The categories boredom, neutrality, indignation, and

[4]Utterances were synthesized from a phonetic description of each of the two previously used sentences, using the IPO Text-To-Speech system (van Rijnsoever, 1988). In this system, two LPC-coded diphone sets were available: one contained about 2000 diphones recorded using the voice of a male speaker (HZ), the other contained about 1600 diphones recorded using another male speaker (PB). The first set of diphones was coded in LPC with 12 poles, the second set with 18 poles. The phonemes were converted into LPC-coded diphones, using both diphone sets. The samples were synthesized at a sampling frequency of 10 kHz. The duration module of the text-to-speech system was kept active, so that the synthetic utterances could be considered adequate neutral expressions. These utterances were then, for each emotion, linearly compressed or expanded using the previously determined mean optimal speech rates (see Table IX). In order to generate monotonous utterances, the intonation module of the text-to-speech system was switched off, yielding a constant F_0. The location of the vowel onsets in the lexically stressed syllables, was determined by listening (Hermes, 1990). An appropriate pitch contour was computed (see Table IV for the corresponding intonation pattern) with the pitch range and the pitch level that were selected as the best ones in the former experiments (see Table VI). Finally, the resulting F_0 curves were imposed on the monotonous synthetic utterances. The voice quality and the fine temporal structure of the diphone-speech were not manipulated.

Table XII. Mean proportion of correct responses for the pitch curves based on optimal pitch level, pitch range, and speech rate applied to synthetic speech

Results of Experiment 7								
	Responses of subjects							
	Neutrality	Joy	Boredom	Anger	Sadness	Fear	Indignation	Mean
Diphone set 1	0.73	0.73	0.90	0.42	0.40	0.35	0.69	0.60
Diphone set 2	0.92	0.50	0.98	0.60	0.54	0.46	0.67	0.67
Overall results	0.83	0.62	0.94	0.51	0.47	0.41	0.68	0.63
Results of other experiments								
Results of Exp. 6	0.80	0.38	0.90	0.38	0.20	0.23	0.50	0.48
Reference results of Exp. 1	0.95	0.73	0.95	0.43	0.93	0.33	0.88	0.74

Table XIII. Confusion matrix for Experiment 7: rule-based generation of emotions from speech synthesized by diphone concatenation

Intended	Responses of subjects							
Emotion	Neutrality	Joy	Boredom	Anger	Sadness	Fear	Indignation	Total
Neutrality	**79**	0	7	7	2	0	1	96
Joy	5	**59**	0	7	0	12	13	96
Boredom	1	0	**90**	2	2	0	1	96
Anger	24	8	1	**49**	0	5	9	96
Sadness	8	2	38	1	**45**	2	0	96
Fear	4	16	0	4	12	**39**	21	96
Indignation	0	2	1	6	12	10	**65**	96

Results are pooled across two presentations of the two diphone sets on the two sentences to twelve subjects.

joy were identified without problem. Fear and anger were better identified in the present experiment than in Experiment 6: 'testing optimal values for pitch level, pitch range, and speech rate on manipulated re-synthesized neutral speech' and even better than in Experiment 1: 'identification of the emotions in the natural utterances'. The possibility that the speaker himself was not really efficient in expressing anger and fear has to be considered.

The most important result, however, is that the rather basic prosodic rules found in this study allow the generation of identifiable emotions in synthetic speech when applied to a conventional synthesizer by diphone concatenation. Another interesting point is that the emotions in Experiment 7, involving synthetic speech, were better identified than in Experiment 6, involving re-synthesized speech, containing the voice quality and micro-duration structure of the natural utterance intended to express neutrality. The difference could be due to the fact that the voice source and the temporal fine structure of neutrality can negatively influence the identification of the other emotions, which is in agreement with the findings of Carlson et al. (1992). This suggests that these supplementary parameters are also of relevance for the vocal communication of emotion.

5. Discussion

In the present study, most of the intended emotions were recognized far above chance level by the subjects. Siegwart and Scherer (1995) report that, in studies where the subjects' task is to infer the underlying emotion only by listening to natural speech, the accuracy of identification of the emotions was found to be approximately 50 or 60%, which is about five times higher than the chance level of 10 or 11%. This corresponds well with the identification rate of 74% for a chance level of 14%, found with the original natural speech in our frame of reference (Experiment 1: 'identification of the emotions in the natural utterances'). Identification rates also remained well above chance level in Experiment 4: 'test of the optimal pitch curves' (58%), in Experiment 6: 'testing optimal values for pitch level, pitch range, and speech rate on manipulated re-synthesized neutral speech' (48%), and in Experiment 7: 'rule-based generation of emotions from speech synthesized by diphone concatenation' (63%). In comparison with the findings of Siegwart and Scherer (1995), the acoustic cues used in this study, i.e., the F_0 level, the F_0 range, the mean speech rate, in combination with controlling the type of pitch contour used, seemed to allow the listeners to reliably identify most emotions. Values were found that are considered to be perceptually optimal for conveying emotion with the parameters concerning pitch level, pitch range, and speech rate. Despite the fact that an eventual partial contribution of the PSOLA-manipulations to the identification of emotions cannot be excluded, such an effect seems quite unlikely, as the range of the manipulations was kept such that it did not introduce clearly conspicuous distortions in the speech signals.

Some emotions appeared to be more difficult to identify than others. For instance, fear was clearly less well identified than boredom, in all experiments. Various explanations for these differences in score can be proposed. In some occasions, the emotion was even poorly identified in the original utterances produced by the speaker, which indicates that some of the original utterances might be, despite the selection of speech material, sub-optimal realizations of the corresponding emotion. In such cases, the values found optimal on a perceptual basis, lead to an increase in correct identification of the emotion. Another explanation is that some emotions may intrinsically be confused more easily with each other, than other emotions, at least on the basis of the speech signal only. This is especially true with speech which excludes elements such as sighs, smacking sounds, or disfluencies, as is the case in the present study. Furthermore, cues other than the ones studied here might be especially relevant for some emotions; voice quality in particular may be a relevant cue (e.g., Ladd et al. (1985); Cummings & Clements, 1995; Laukkanen et al., 1997).

The present findings seem to support the idea that emotions are signaled by a complex interaction of prosodic cues which, in principle, can be controlled in synthesized speech (e.g., Murray & Arnott, 1993). Despite some limitations of the present study, i.e., a limited number of emotions, only two sentences, a single

speaker, relatively few listeners per experiment, and a limited number of acoustic parameters, the results obtained are quite encouraging. Quantitative values for a possible modeling of seven emotions have been proposed, that are based on, and have been evaluated through, perceptual measurements. Because of their relative simplicity, the parameters can, in principle, be manipulated in a broad range of synthesizers for the purpose of generating emotional speech. Although it was not tested in the present study, these values could be used in the field of speech recognition for identifying the speaker's emotion. Consequently, these values allow a 'speech module' to perform synthesis as well as recognition of emotion in speech, which will be most useful in combination with a user model taking the emotion of the user into account. Such a user model of emotion-cognition, in which, among others, prosodic elements are considered for determining the emotional state of the user was proposed by Lisetti (1999). Bianchi-Berthouse and Lisetti (in this issue) also propose a Model Of User Emotion making use, among others, of auditory input. However, it has to be kept in mind that a unique modeling of emotion is probably not possible. One knows from everyday life experiences, that a particular emotion can be conveyed in many ways, depending on the situation and on the cultural context. Anger, for example, can be expressed openly, freeing one's mind, or can be more or less repressed. Besides, individual speakers can prefer the use of different acoustic parameters to communicate the same emotion. The expression of emotions can also appear as more or less convincing, appropriate, trustworthy, and intense (Bartneck, in this issue). Furthermore, as reported by Carlson et al. (1992), extra factors such as sighs, linguo-dental smacks, voice breaks, and jitter can also contribute decisively to the vocal expression of emotion.

An important concern is the possible generalization of the present findings. In the experimental design, precautions were taken in order to maximize validity, and subjects' answers appeared to be quite consistent. People seem to agree quite well, at least for Dutch, on how the expression of specific emotions sounds. It does not seem unreasonable to generalize the perceptual behavior of our subjects to typical behavior of Dutch listeners. Moreover, there is no reason to expect that the results could not be generalized to experimental settings in which other intonation grammars and other types of synthesizer would be used.

Part of the same validity issue is the potential difference between elicited and spontaneous emotions. As a first step in research of emotional speech, the advantages of high-quality recordings, with control over the acoustic environment and over the content of utterances offered by elicited speech, are obvious. Although Williams and Stevens (1972), who compared acted and spontaneous emotional speech, concluded that data obtained in spontaneous speech are not inconsistent with data obtained from acted emotions, further study involving spontaneous speech will become more attractive once more detailed rules have been determined. The same argument holds for potential differences in expression of emotion among different speakers. The fact that a single speaker was used in this study forms a limitation of the investigation. However, Mozziconacci (1998) showed that various speakers

appeared to make a fairly consistent use of speech parameters in the expression of emotion. Speakers expressing a given emotion seem to have personal preferences for particular cues on which they rely more than on other useful cues. It does not necessarily imply discrepancies among speakers, at least not if the data is considered in a qualitative way, rank-ordering the emotions according to increasing/decreasing parameter values.

Furthermore, although the focus of the present study was restricted to acoustic parameters conveying emotion in speech, the interpretation of the whole spoken message by the listener involves much more than prosody alone. The inference of meaning occurs in a specific situational context, in a given language, between people of specific personalities, gender, cultural and educational backgrounds. It involves a particular semantic content, prosodic variations, as well as correspondence or mismatch between the previous elements of the communication. The prosodic information may therefore constitute a useful part of the information needed for an approach in the larger framework of pragmatics.

The parameter settings found to be optimal for conveying the emotions considered in the present study, are reported on the left hand side of Table XIV, averaged over the two sentences, while they were given separately for each sentence in Table VI, with the intonation patterns. To facilitate comparison with the results of related studies, the mean pitch and the mean distance between initial and final frequency, referred to as 'declination range', were calculated for the speech material of Experiment 7. The values averaged over both sentences and both diphone sets are reported on the right hand side of Table XIV. These measures are convenient because they conform to those in related studies. From the two values for pitch range: excursion size of the pitch movements given in column 2 and 'declination range' given in

Table XIV. Parameters found optimal in the present study (in the left hand columns) and calculations on the speech material of Experiment 7: rule-based generation of emotions from synthetic speech averaged over two sentences and two diphone sets (in the right hand columns)

	Optimal values			Calculations on speech material of Experiment 7	
Expression	End frequency (Hz)	Excursion size of pitch movements (s.t.)	Duration relative to neutrality (%)	Mean F_0 (Hz)	'Declination. range'* (s.t.)
Neutrality	65	5	100	94.0	6.5
Joy	155	10	83	246.2	6.1
Boredom	65	4	150	93.8	8.0
Anger	110	10	79	185.6	6.0
Sadness	102	7	129	160.3	7.3
Fear	200	8	89	277.0	6.5
Indignation	170	10	117	279.0	7.1

*'Declination range' is here defined as the mean distance between begin and end frequency.

column 5, the highest of the two indicates the largest variation range, and is taken as a basis for comparison with other studies.

Although comparison between the results of different studies remains a difficult task, as long as no standards are widely used, a rather global comparison can be made with a few related studies. It shows that the results of the present study correspond quite well with results from these related studies. Indeed, a clear resemblance can be observed with findings for Dutch from van Bezooijen (1984); except for the pitch level in the expression of fear, the pitch level and pitch range values are similar in both studies. The pitch levels which Fairbanks and Pronovost (1939) reported match quite well the ones reported here, but they describe much wider pitch ranges. Although the pitch levels reported by Carlson et al. (1992) are in a narrower and lower range compared with the ones reported here for the corresponding emotions, the ranking of the emotions from lower to higher pitch is the same as in the present study, except for neutrality. The findings of Kitahara and Tohkura (1992) also seem to be in agreement with the present ones. Present results are thus not specific to the present approach, which considers the perception of emotion in correspondence with its production in natural speech, and was carried out within the framework of an intonation model.

In summary, it can be concluded that this study represents a step towards understanding how emotion is conveyed through prosody. The modeling proposed here seems, at the moment, to be quite acceptable for male speech in Dutch. Furthermore, the methodological approach not only made it possible to qualify the differences between emotions, but also to quantify them in terms of deviations from the parameter values found in neutral settings. The resulting rules may be used in synthesis of emotionally colored speech as well as recognition of speaker's emotions from the acoustic signal (Mozziconacci, 1998). These values found optimal for the perception of emotion in speech can best be used in combination with a user model including a description of the emotional state of the user, which would result in allowing user adaptation in computer systems by taking the emotion of the user into account.

Finally, the results of the present study more or less corroborate results of related studies of speech conveying emotion. However, in contrast with these previous studies, it relates speech production data to speech perception data, and successively involves natural speech, re-synthesis of speech allowing manipulations of natural speech, and rule-based synthesis of speech. Additionally, it involves the use of a methodological framework for the investigation of intonation. The intonation model used in this framework is perception-oriented, and performs a data reduction by simplification of the pitch contour on the basis of perception. In this way, perception functions as a filter, as it seems desirable to model only the information that is perceptually relevant to the communication. Hence, this paper proposes to model the listener, considering that both the machine and the user can be the listener, depending on the system. Indeed, results may be used in Text-To-Speech systems as well as in systems for speech recognition.

Acknowledgement

I would like to sincerely thank Dik Hermes for his valuable comments on a draft version of the present paper.

References

Bartneck, C.: How convincing is Mr. Data's smile: Affective expressions of machines (in this issue).

Beckman, M. E.: 1997, Speech models and speech synthesis. In: J. P. H. van Santen, R. W. Sproat, J. P. Olive and J. Hirschberg (eds.) *Progress in speech synthesis*. Springer-Verlag, New York, pp. 185–209.

Bezooijen, R. A. M. G. van: 1984, *The characteristics and recognizability of vocal expression of emotion.* Foris, Dordrecht, The Netherlands.

Bianchi-Berthouse, N. and Lisetti C. L.: Modeling multimodal expression of user's affective subjective experience (in this issue).

Bouwhuis, D. G.: 1974, The recognition of attitudes in speech. *IPO Annual Progress Report* **9**, pp. 82–86.

Cahn, J. E.: 1990, Generating expression in synthesized speech. *Technical report*, MIT Media Lab., Boston.

Carlson, R.: 1991, Synthesis: modelling variability and constraints. *Proceedings Eurospeech'91*, Genova, Italy **3**, pp. 1043–1048.

Carlson, R., Granström, B., and Nord, L.: 1992, Experiments with emotive speech: acted utterances and synthesized replicas. *Proceedings ICSLP 92*. Banff, Alberta, Canada, **1**, pp. 671–674.

Charpentier, F., and Moulines, E.: 1989, Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Proceedings Eurospeech'89*. Paris, France, **2**, pp. 13–19.

Collier, R.: 1991, Multi-language intonation synthesis. *Journal of Phonetics* **19**, pp. 61–73.

Cosmides, L.: 1983, Invariances in the acoustic expression of emotion during speech. *Journal of Experimental Psychology: Human Perception and Performance* **9**, pp. 864–881.

Cummings, K. E., and Clements, M. A.: 1995, Analysis of the glottal excitation of emotionally styled and stressed speech. *Journal of the Acoustical Society of America* **98**(1), pp. 88–98.

Ekman, P.: 1982, *Emotion in the human face, second edition.* Cambridge University Press, New York.

Fairbanks, G. and Pronovost, W.: 1939, An experimental study of the pitch characteristics of the voice during the expression of emotion. *Speech Monographs* **6**, pp. 87–104.

Frick, R. W.: 1985, Communicating emotion: the role of prosodic features. *Psychological Bulletin* **97**, pp. 412–429.

Frijda, N. H.: 1986, *The emotions*. Cambridge University Press, Cambridge, England.

Hart, J. 't, Collier, R. and Cohen, A.: 1990, *A perceptual study of intonation.* Cambridge University Press, Cambridge.

Hermes, D. J.: 1988, Measurement of pitch by subharmonic summation. *Journal of the Acoustical Society of America* **83**, pp. 257–264.

Hermes, D. J.: 1990, 'Vowel-onset detection. *Journal of the Acoustical Society of America* **87**(2), pp. 866–873.

House, D.: 1990, *Tonal perception in speech.* Lund University Press, Lund.

Izard, C. E.: 1977, *Human emotions*. Plenum Press, New York.

Kitahara, Y. and Tohkura, Y.: 1992, Prosodic control to express emotions for man-machine interaction. *IEICE Transactions on Fundamentals of Electronics, communications and computer sciences* **75**, pp. 155–163.

Ladd, D. R., Silverman, K. E. A., Tolkmitt, F., Bergman, G. and Scherer, K. R.: 1985, 'Evidence for the independent function of intonation contour type, voice quality, and F_0 range in signalling speaker affect. *Journal of the Acoustical Society of America* **78**, pp. 435–444.

Laukkanen, A.-M., Vilkman, E., Alku, P. and Oksanen H.: 1997, On the perception of emotions in speech: the role of voice quality. *Journal of Logopedics and Phoniatrics Vocology* **22**(4), pp. 157–168.

Leinonen, L., Hiltunen, T., Linnankoski, I. and Laakso, M.-L.: 1997, 'Expression of emotional-motivational connotations with a one-word utterance. *Journal of the Acoustical Society of America* **102**(3), pp. 1853–1863.

Lieberman, P. and Michaels, S. B.: 1962, Some aspects of fundamental frequency and envelope amplitude as related to emotional content of speech. *Journal of the Acoustical Society of America* **34**, pp. 922–927.

Lisetti, C. L.: 1999, A user model of emotion-cognition. *Proceedings of the workshop on attitude, personality, and emotions in user-adapted interaction, at the 7th International Conference on User Modeling (UM'99)*. Banff, Canada.

Mozziconacci, S. J. L.: 1998, *Speech variability and emotion: Production and perception.* Technical University Eindhoven, The Netherlands.

Mozziconacci, S. J. L., and Hermes, D. J.: 1999, Role of intonation patterns in conveying emotion in speech. *Proceedings ICPhS 99*. San Francisco, USA.

Murray, I. R.: 1989, *Simulating emotion in synthetic speech.* University of Dundee, Scotland, UK.

Murray, I. R. and Arnott, J. L.: 1993, Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion. *Journal of the Acoustical Society of America* **93**, pp. 1097–1108.

Pijper, J.-R. de: 1983, *Modelling British English intonation: an analysis by resynthesis of British English intonation*. Foris, Dordrecht, The Netherlands.

Plutchik, R.: 1980*, Emotion: a psychoevolutionary synthesis.* Harper & Row, New York.

Protopapas, A. and Lieberman, P.: 1997, Fundamental frequency of phonation and perceived emotional stress. *Journal of the Acoustical Society of America* **101**(4), pp. 2267–2277.

Rijnsoever, P. van: 1988, A multilingual text-to-speech system. *IPO Annual Progress Report* **23**, 34–39.

Rosis, F. de, and Grasso, F.: in press, Affective natural language generation. In: A. Paiva (ed.): *Affect in interaction*. Springer LNAI Series, in press.

Siegwart, H. and Scherer, K. R.: 1995, Acoustic concomitants of emotional expression in operatic singing: the case of Lucia in Ardi gli incensi. *Journal of Voice* **9**(3), pp. 249–260.

Verhelst, W. and Borger, M.: 1991, Intra-speaker transplantation of speech characteristics: an application of waveform vocoding techniques and DTW *Proceedings Eurospeech'91*. Genova, Italy**, 3**, pp. 1319–1322.

Williams, C. E. and Stevens, K. N.: 1972, Emotions and speech: some acoustical factors. *Journal of the Acoustical Society of America* **52**, pp. 1238–1250.

Zelle, H. W., Pijper, J.-R. de and Hart, J. 't: 1984, Semi-automatic synthesis of intonation for Dutch and British English. *Proceedings Xth ICPhS*. Utrecht, The Netherlands, IIB, pp. 247–251.