Prosodic predictors of upcoming positive or negative content in spoken messages

Marc Swerts[a)]
Communication and Cognition, Tilburg University, P.O. Box 90153, NL-5000 LE Tilburg, The Netherlands

Julia Hirschberg
Department of Computer Science, Columbia University, 1214 Amsterdam Avenue, New York, New York 10027

(Received 19 March 2010; revised 28 June 2010; accepted 29 June 2010)

This article examines potential prosodic predictors of emotional speech in utterances perceived as conveying that good or bad news is about to be delivered. Speakers were asked to call an experimental confederate to inform her about whether or not she had been given a job she had applied for. A perception study was then performed in which initial fragments of the recorded utterances, not containing any explicit lexical cues to emotional content, were presented to listeners who had to rate whether good or bad news would follow the utterance. The utterances were then examined to discover acoustic and prosodic features that distinguished between good and bad news. It was found that speakers in the production study were not simply reflecting their own positive or negative mood during the experiment, but rather appeared to be influenced by the valence of the positive or negative message they were preparing to deliver. Positive and negative utterances appeared to be judged differently with respect to a number of perceived attributes of the speakers' voices (like sounding hesitant or nervous). These attributes correlated with a number of automatically obtained acoustic features.
 [DOI: 10.1121/1.3466875]

PACS number(s): 43.70.Fq, 43.71.Sy, 43.70.Mn, 43.71.Gv [SSN] Pages: 1337–1345

I. INTRODUCTION

Listeners sometimes have the sensation, when being addressed by another person, that the speaker's voice betrays that he or she will come up with good or bad news. For instance, even though the lexical content in the opening sequence of a telephone chat may not seem to reveal anything crucial, one somehow can tell from the way the person at the other end starts talking or says "Hello" what the emotional content will be of the subsequent utterances. This impression relates to a more general problem which has received some scholarly attention, i.e., the extent to which prosodic features may *presignal* elements of information that occur later in a speaker's utterance or discourse, and whether such features have cue value to listeners who process incoming speech.

The question of prosodic prediction is intriguing in view of many speech production models, especially those inspired by Levelt's model (Levelt, 1989). Those models often make specific claims about spontaneous speech production, and assume that speakers have access to limited look-ahead, typically no more than a clause. Therefore, evidence of possible signaling of upcoming content is challenging for such models, particularly when it implies relatively large amounts of cognitive preplanning. Interestingly, recent studies do show that speakers exploit prosodic features to presignal linguistic content that occurs later in the discourse, as we will discuss in Sec. II. While such findings do not necessarily falsify claims made by current speech production models, it remains to be seen how they can be incorporated in models that assume only limited amounts of preplanning. In any case, studies so far have dealt with predictive cues to non-emotional features. To date, we lack data on the extent to which prosody can also be used to signal upcoming emotional or attitudinal state, even while much other research suggests that emotions **do** have acoustic correlates in the speech signal.

In this study we examine whether prosody may also serve as predictor of upcoming emotional content. In Sec. II we discuss previous work on prosodic prediction and prosodic correlates of emotion. In Sec. III we present a production study eliciting utterances with a negative or positive content in a spontaneous, but controlled manner. In Sec. IV we describe a perception study in which observers are asked to judge the content of upcoming information ("good news" or "bad news") under two different conditions, i.e., in utterances presented as audio files or as text only. In Sec. V we examine the prosodic correlates of decisions made in our perception study. We discuss our results and conclude in Sec. VI.

II. PREVIOUS WORK

A. Prosodic prediction

Previous studies have shown that prosodic features can presignal elements of information occurring later in an utterance. Different researchers have found that prosodic features can be used to presignal an upcoming linguistic boundary, such as the end of a sentence or a speaking turn in a conversation. It has been claimed that this phenomenon may explain why the turn-taking mechanism typically proceeds very fluently (Levinson, 1983; Koiso *et al.*, 1998; Ward and Tsu-

a)Author to whom correspondence should be addressed. Electronic mail: m.g.j.swerts@uvt.nl

kahara, 2000). These cues may be local in nature, including specific boundary tones or lengthened syllables (Carlson *et al.*, 2005; Swerts *et al.*, 1994; Barkhuysen *et al.*, 2008). Other research claims that the more global declination pattern, defined as the gradual lowering of pitch in the course of an utterance, can also be used as a cue to the end of a linguistic unit, given that the final low tone of a speaker appears to be rather constant (Hart *et al.*, 1990; Leroy, 1984; Grosjean, 1983). Note, however, that these latter studies have been criticized as identifying artifacts of read-aloud sentences where the lay-out or the visual feedback gives clues about the size of the speech unit (Umeda, 1982).

Other possible predictors of later information are *filled pauses* ("uh" or "uhm"). These may signal speakers' word search problems in their mental lexicon, so that the filled pauses can have a beneficial effect on listeners ability to recognize words in upcoming speech (Fox Tree, 2001). Filled pauses may also occur at the beginning of speaker responses that are uncertain (Smith and Clark, 1993, Brennan and Williams, 1995; Swerts and Krahmer, 2005), or before important discourse boundaries (Swerts, 1998). Finally, filled pauses have been claimed to function as turn-holding or turn-taking cues that make it clear to addressees that the interlocutor either wants to keep or to take the turn (Gravano and Hirschberg, 2009). Related to filled pauses, it appears that speaker uncertainty is conveyed not just on the word representing the uncertain content, but also through prosodic features in the preceding context (Liscombe *et al.*, 2005; Pon-Barry, 2008; Pon-Barry and Shieber, 2009).

Other work suggests that prominence patterns, both at the level of the word and at higher linguistic units, may presignal elements of linguistic structure that occur later. Some studies have shown that the presence or absence of (lexical) stress on the initial syllable of a word may enable a listener to select a word from the mental lexicon before it is fully uttered (Cutler *et al.*, 1997). At higher levels, it has been shown that pitch accents, especially those occurring in prenuclear position, may presignal an upcoming contrast. For instance, an utterance which begins with "I do not want the BLUE ball,…," strongly suggests that the speaker will introduce a ball of a different color in the upcoming clause (Swerts, 2007). Also, there are claims in the literature that nuclear accents in prosodic phrases with a specific shape (like an early timed pitch fall) make it clear to the listener that this is the final accent in a prosodic phrase, and that no other accent will follow (Silverman and Pierrehumbert, 1990; Ladd, 1996; Krahmer and Swerts, 2001). Similarly, studies conducted within the visual world paradigm have shown that pitch accents can redirect eyegaze patterns of listeners before they have heard the complete word or utterance (Dahan *et al.*, 2002).

B. Correlates of emotion

While some studies have shown that prosodic information may signal upcoming phrase boundaries, lexical access or discourse structure, no one has previously examined prosodic features that presignal attitudinal or emotional content. However, there has been a great deal of work on how acoustic and prosodic features correlate with emotions and attitudes. Past studies have examined the way emotions and attitudes are displayed in specific facial expressions (Ekman and Friesen, 1975) as well as intonational and rhythmic patterns and variation in voice quality (Bänziger and Scherer, 2005). Research to date has found that the relation between emotions and acoustic or visual features is a complex one. Mozziconacci (1998) showed that, in the auditory domain, listeners could reliably distinguish between basic sets of emotions on the basis of systematically varied prosodic variables. Computational studies have found that the 'classic' emotions such as anger, happiness, and sadness can be automatically classified from acoustic and prosodic features with fair accuracy (Batliner *et al.*, 2006; Yuan *et al.*, 2002; Ang *et al.*, 2002; Bitouk *et al.*, 2010). Other research has focused on valency, i.e., the extent to which an emotion is perceived as positive or negative (Truong, 2009). In a previous study on professional newsreaders on Dutch public TV (Swerts and Krahmer, 2010), it was found that audiences can reliably guess, on the basis of the expressive style of newsreaders, whether they are covering a positive or negative news item. In that study, it was found that positive markers show more expressive features than negative markers. Other computational work has reported distinguishing positive from negative emotions in a call-center corpus (Lee and Narayanan, 2005). A comprehensive overview of current research into automatic classification of emotions is given by Schuller *et al.* (2009).

An important limitation in many studies of emotional speech is the fact that the data studied may not be representative of data in natural settings. Many studies have collected data in experimental settings, where speakers are explicitly instructed to portray specific sets of emotions (Scherer, 2003). Unfortunately, there is evidence that these kinds of "acted" emotions may not constitute a good basis for models about human signals to emotion, as subjects' display in these circumstances is more stereotypical and more exaggerated than that observed from spontaneous speakers (Wilting *et al.*, 2006). Therefore, there is a growing awareness that better elicitation procedures are needed to induce emotions naturally in subjects, so to produce more realistic data. Conversely, corpus-based studies have their limitations as well. One major stumbling block is the extent to which emotions can reliably be annotated in spoken databases. Not only is it is difficult to reach consensus among multiple labelers, but it is hard to establish whether the annotated emotions were indeed those experienced by the original speakers. In other words: in this particular area of emotion research, there appears to be a need for data that are representative of natural speech productions, and at the same time allow a straightforward specification of the produced emotions (Scherer, 2003).

III. PRODUCTION STUDY

Our production study was designed to elicit utterances with a negative or positive content through a semi-spontaneous procedure. Participants were asked to imagine that they would have to inform a job applicant whether or not she or he would be given the job. The elicitation procedure

included a form of mood induction to produce a positive or negative mood in each participant before they performed the task.

A. Procedure

Participants in the production study were told that the goal of the experiment was to learn more about the effect of mood on characteristics of their language and that their voices would be recorded. The next step in the experiment was to assess subjects' current mood and to attempt to induce either a positive or negative mood via *mood induction*. Subjects' current mood was first assessed via a short questionnaire, taken from Krahmer *et al.* (2004). The questionnaire consisted of 6 bipolar 7-point scales (Dutch), whose items can be translated into English as follows: happy/sad, pleasant/unpleasant, satisfied/unsatisfied, content/discontent, cheerful/sullen and high spirits/low-spirited. For 4 scales, "1" corresponded to a very positive mood and "7" to a very negative one, while for the other 2 scales, the extremes of the scales were reversed; this variation was introduced to increase subjects' attention to the task. Next, to induce a positive or a negative mood, subjects were asked to watch one of two movie fragments. In the positive mood-induction condition, participants viewed the first 7 min (excluding the initial tune) of Episode 5.14 of *Friends. The one where everyone finds out*. Participants in the negative mood-induction condition watched a 7-min fragment of Schindler's list (the "Liquidation of the ghetto, March 1943"-scene), corresponding to scene 13/14 of Disk 1 of the commercial DVD. The former is generally considered to be a very funny part of this mainstream show, whereas the latter contains a quite depressing and dark scene of a movie on the Holocaust. After watching one of these two film clips, participants were again asked to complete the same mood-evaluating questionnaire to assess the success of mood induction.

Subjects were then asked to inform a female job candidate for an open position at the university that she had been successful or unsuccessful in getting a job by leaving a message in the job candidate's voicemail. The reason for eliciting data through voice mail rather than through an interaction with a confederate was to minimize the effect of the confederate's interaction on the performance of the subjects. The scenario of a job interview was chosen to present a plausible emotional situation. People in whom the negative mood was induced were to leave the 'sorry no job' voicemail and people who had received positive mood induction were to leave the 'you got the job' message. There were 2 conditions in both the positive and the negative tasks. In one, the message left in voice mail actually contained the positive or negative information about the decision; in the other, the applicant was simply told that the committee had made a decision, and that she was to call back to discover the outcome. The design was between-subject, as participants only participated in one of the 4 conditions: positive with decision included in message, negative with decision included, positive without decision included and negative without decision. Following the recording session, participants were again asked to complete the mood questionnaire with the 6 scales.

B. Participants

In total, 40 speakers (25 female, 15 male) volunteered to participate in the elicitation experiment. Participants were all students from the University of Tilburg (The Netherlands), native speakers of Dutch, who participated as a partial fulfillment for course credit. They were equally distributed over the 4 conditions of the study.

C. Data

Table I gives some examples of Dutch utterances (with their English translations) to illustrate the kinds of elicitations in the different conditions of the production study. They represent cases of positive and negative calls from both conditions with and without decision included; the telephone number and the names in the examples (and in the actual experiment as well) are fake. The examples for the 'with decision' data are typical in that they consist of an introductory part, in which the speaker explains who he or she is and provides some context, after which the positive or negative decision about the job application is revealed. The latter part is presented between square brackets in the examples in the table, and, as will be explained below, is the part of the message that will be isolated from its context in further perception experiments described in Sec. IV. The examples for the 'without decision' data show that the negative and positive mood condition elicit very comparable utterances, whose lexical contents do not reveal any obvious cues to the emotional content of the decision the speaker wants to convey to the addressee.

D. Results and discussion

Table II shows results of different stages of the mood induction procedure for subjects in the positive and negative conditions, leaving voicemail with and without providing a decision.

A reliability analysis showed that the scores for the 6 different 7-point scales were very consistent with a Cronbach's α of 0.923. Therefore, for further analyses the scores for the different scales were averaged; these average scores are presented in the table. The data were analyzed by means of a repeated measurements analysis of variance with stage (beforeMI, afterMI, after-Exp) as a within-subject factor, with mood condition (positive, negative) and experimental condition (with decision, without decision) as between-subject factors, and with the average mood scores as the dependent variable. The analysis revealed a main effect of mood ($F_{(1,37)}=15.102$; $p<001$; $\eta_p^2=0.290$) with the negative mood-induction overall leading to more negative scores (5.64) than the positive mood induction (4.79). There was also a main effect of stage ($F_{(2,74)}=11.539$; $p<001$; $\eta_p^2=0.238$); post-hoc bonferroni tests showed that the afterMI moment differed significantly (4.99) from the beforeMI (5.40) and afterExp stages (5.26). There was also a significant 2-way interaction between mood and stage ($F_{(2,74)}=31.306$; $p<001$; $\eta_p^2=0.458$), which can be explained as follows: before the experiment starts, the reported moods of participants are basically the same (negative: 5.45; positive:

TABLE I. Examples of original Dutch utterances (with English translations within brackets) elicited in positive and negative mood conditions of the 'with decision' and 'without decision' experiments.

Experiment	Mood	Example
With decision	Positive	"Goeie morgen, Mirjam, u spreekt met Mieke Peeters van L&L. We hebben vorige week sollicitatiegesprekken gevoerd, en ik wil je mededelen dat [je bent aangenomen, en ik hoop dat je nog steeds geïnteresseerd bent in deze functie. Zou je mij zo snel terug kunnen bellen op 03/22234476. Alvast bedankt. Groetjes.]" *("Good morning, Mirjam, this is Mieke Peeters talking from L&L. We have had interviews last week, and I would like to inform you that [we want to offer you the job, and I hope that you are still interested in this position. Would you be so kind to call me back at 03/22234476 as soon as possible. Thanks a lot. Cheers.]")*
	Negative	"Goeie middag, Mirjam, je spreekt met Michael Dooren van L&L. We hebben laatst een gesprek gehad, een sollicitatiegesprek gehad. En ik wou heel even doorgeven dat [je voor ons niet geschikt genoeg bent bevonden voor deze functie. Als je vragen hebt, kun je nog contact met ons opnemen. Hartstikke bedankt. Tot ziens.]" *("Good afternoon, Mirjam, this is Michael Dooren from L&L. We have recently had a talk, we had an interview. And I would like to let you know that [we feel you are not suitable enough for this job If you have any questions, then you can contact us. Thanks a lot. See you.]")*
Without decision	Positive	"Goeie middag, Mirjam, met Gaby van L&L, de personeelsmanager. We hebben de uitslag van het sollicitatiegesprek. En je zou me terug kunnen bellen op 03/22234476. Als je hiervoor de uitslag wilt hebben, moet je mij bellen. Ik herhaal nog een keer: 03/22234476. Fijne dag nog verder. Daag." *("Good afternoon, Mirjam, this is Gaby from L&L, the human resources manager. We know the outcome of the interview. And you can call me back at 03/22234476. So if you want to know the result, you have to call me back. I repeat once again: 03/22234476. Have a nice day. Bye.")*
	Negative	"Hoi, je spreekt met Rianne, namens L&L. We hebben vorige week een sollicitatiegesprek met jou gehad, en de uitslag daarvan is bekend. Als je deze uitslag wil weten, kan je bellen op 03/22234476. Ik herhaal: 03/22234476. Doei!" *("Hi, this is Rianne speaking, on behalf of L&L. We have had job interviews last week with you, and we have reached a conclusion on this. If you would like to be informed about the outcome, you can call us at 03/22234476. I repeat: 03/22234476. Take care!")*

5.34), but the scores become more negative (5.74) or positive (4.24) after participants have seen the corresponding positive or negative movie clip. After having produced their voicemail, participants' scores become less positive or negative (negative: 5.73; positive: 4.79).

Our data collection procedure thus appears to have been successful in inducing the desired mood in subjects who were to leave positive or negative messages. We now turn to the question of whether these induced moods could be interpreted by listeners.

IV. PERCEPTION STUDY

We used the data collected in our production studies as materials for a set of perception studies. We first excised the initial portions of each message—the portion **before** the actual decision was revealed in the 'with decision' condition and the entire message in the 'without decision' condition. As a result, the stimuli for the 'without decision' experiment were generally longer, as they consisted of the entire message the speakers had left on the answering machine. In the former case, this would correspond to the portion of messages presented within square brackets in Table I.

We then examined these initial messages to determine whether any contained lexical cues to the upcoming good or bad news. While we did not find such cues, we nonetheless presented both text-only and speech versions of these stimuli to listeners to verify our observation. We presented these initial messages to listeners in the two modality conditions and asked them to judge whether the job candidate would be offered the job (good news) or not (bad news).

TABLE II. Average mood induction scores with standard deviations (with higher numbers representing more negative mood) at three different stages in experiments with and without decision included: Before Mood induction (BeforeMI), immediately after Mood induction (AfterMI), and immediately after the production experiment (AfterExp).

		Mood	
Experiment	Stage	Positive	Negative
With decision	BeforeMI	5.48 (0.27)	5.59 (0.25)
	AfterMI	4.22 (0.24)	5.91 (0.22)
	AfterExp	4.72 (0.24)	5.82 (0.23)
Without decision	BeforeMI	5.20 (0.27)	5.35 (0.27)
	AfterMI	4.26 (0.24)	5.58 (0.24)
	AfterExp	4.89 (0.24)	5.67 (0.24)

A. Procedure

Participants took part in one of 2 perception studies, conducted using *wwstim*, a software package used to set-up and conduct rating experiments via the internet. In the speech condition, participants had to rate initial message recordings of all 40 speakers of the production studies described above. They could only listen to the speech, without any written text. To compensate for possible learning effects, all participants were presented with a different random order of the stimuli within a block. The two blocks consisted of all the utterances of either the 'with decision' or 'without decision' conditions. The block with utterances from the 'with

TABLE III. Average proportion of positive scores, standard deviations, and F-statistics for utterances (both speech and text versions) elicited in positive and negative mood conditions in experiments with and without decision included, and difference scores between ratings for speech and text versions.

		Mood				
Experiment	Condition	Positive	Negative	$F_{(1;18)}$	p-value	η_p^2
With decision	Speech (1)	0.60 (0.28)	0.30 (0.21)	7.210	<0.05	0.286
	Text (2)	0.37 (0.17)	0.39 (0.21)	0.021	n.s.	0.001
	Δ-score (1–2)	0.22 (0.30)	−0.09 (0.25)	6.243	<0.05	0.258
Without decision	Speech (1)	0.47 (0.32)	0.51 (0.26)	0.091	n.s.	0.005
	Text (2)	0.35 (0.27)	0.53 (0.32)	2.090	n.s.	0.104
	Δ-score (1–2)	0.12 (0.42)	−0.01 (0.27)	0.716	n.s.	0.038

decision' experiment were always presented first. The instruction given to the participants was to rate each utterance (by forced-choice) according to whether they thought it would be followed by a negative or positive decision. The text-only experiment was essentially the same as the speech-version, except that raters saw only text transcripts of the speech.

B. Participants

For the text-with-speech experiment, 95 participants were recruited, with 32 (33.7%) men, and 63 (66.3%) women. Their average age was 33.7. The text-alone experiment was conducted with 39 participants. Unfortunately, the demographics of participants in that experiment were not recorded. None of the participants had participated in the production study, and each participated in only one of the two perception experiments.

C. Results and discussion

The results of the two perception studies are given in Table III. This table shows the proportion of positive scores for positive and negative utterances (audio and text-alone conditions), for cases where the actual decision was subsequently included or not. The data were analyzed with a number of repeated measurements analyses of variances, always with mood as independent factor (positive or negative). Table III reveals a number of effects. For the condition where the negative or positive decision was included in the actual message (even when that part was excised from the stimuli presented to listeners), the perception data reveal a significant difference between judgments for the positive and negative conditions. Note that the text-only conditions show no such difference, confirming our observation that there were no lexical cues in these initial portions of the message: the difference scores comparing ratings of the text and speech conditions are also significantly different. In the speech version ratings are more positive in the positive context, and more negative in the negative condition. Interestingly, in the 'without decision' condition, when the complete message did not contain the actual positive or negative decision, there appear to be no such effects, for either the speech or text versions of the experiment. The modality difference scores (speech vs text) are not significant either for the 'without decision' condition, though it can be observed that the scores show the same trends as for the other condition.

The perception experiment thus confirms our determination that lexico-syntactic features do not have a significant cue value in the stimuli presented to judges, as the text-only version of the experiment does not reveal significant results, whereas listeners could significantly judge the upcoming emotional content when they had access to the speech. This main finding suggests that speakers are attending to acoustic and prosodic features in the messages to identify the valence of the upcoming content. In addition, it is most interesting to note that neither the ratings of speech nor the text messages reveal significant effects for the condition in which the actual decision about the job was **not** in the full message, despite the fact that the emotional self-reports of speakers after the initial mood induction were essentially the same between the two production experiments (with and without decision). This difference suggests that speakers in the production study were not simply reflecting their own positive or negative mood in the 'without decision' experiment, since these moods were not significantly different between the two conditions. It seems rather to indicate that speaker productions are influenced by the valence of the actual message they are preparing to deliver in the 'with decision' condition, whereas in the 'without decision' condition they have no such task before them, even though they know the valence of the decision. In other words: it may be the case that the utterances in the 'with decision' experiment only reveal that the speaker is feeling uneasy about the unpleasant news he or she needs to convey.

V. MEASUREMENTS OF PROSODIC CORRELATES

Given our findings above that listeners in the perception study could detect whether a speaker is going to introduce good or bad news in the later part of a message from only a lexically neutral initial message, and given that only the speech condition revealed these differences, we now examine what information may have led raters to their decision. We first look at additional subjective judgments of emotional content and then at potential acoustic and prosodic cues of these judgments and of the original valence judgments.

TABLE IV. Average perceived voice attributes and standard deviations for utterances elicited in positive and negative mood conditions in experiments with and without decision included, and respective *T*-statistics.

		Mood			
Experiment	Judgment	Positive	Negative	*t*-stats	*p*-value
With decision	Friendly	3.62 (0.55)	3.23 (0.53)	1.603	n.s.
	Hesitant	2.60 (0.76)	3.15 (0.68)	−1.695	n.s.
	Concerned	2.31 (0.49)	2.78 (0.32)	−2.550	<0.05
	Smiling	2.57 (0.83)	1.91 (0.39)	2.261	<0.05
	Certain	3.27 (0.81)	2.78 (0.37)	1.669	n.s.
	Nervous	2.12 (0.69)	2.58 (0.38)	−1.854	=0.08
	Sad	1.68 (0.51)	2.31 (0.49)	−2.793	<0.05
	Pleasant	3.23 (0.62)	2.93 (0.46)	1.225	n.s.
	Interested	3.33 (0.57)	2.95 (0.34)	1.795	=0.09
	Monotone	2.22 (0.54)	2.78 (0.61)	−2.179	<0.05
Without decision	Friendly	3.61 (0.52)	3.43 (0.51)	0.737	n.s.
	Hesitant	2.92 (0.95)	2.85 (0.47)	0.183	n.s.
	Concerned	2.56 (0.42)	2.61 (0.26)	−0.296	n.s.
	Smiling	2.21 (0.62)	2.18 (0.59)	0.085	n.s.
	Certain	3.01 (0.86)	3.18 (0.56)	−0.525	n.s.
	Nervous	2.38 (0.79)	2.40 (0.49)	−0.053	n.s.
	Sad	2.21 (0.57)	2.40 (0.51)	−0.792	n.s.
	Pleasant	3.07 (0.67)	2.91 (0.55)	0.584	n.s.
	Interested	3.14 (0.47)	2.94 (0.64)	0.830	n.s.
	Monotone	2.53 (0.79)	2.91 (0.62)	−1.190	n.s.

A. Procedure

The judgment test again made use of the WWSTIM software. The 40 stimulus utterances were identical to those in the previous perception experiment, and were presented to participants in 2 blocks, i.e., utterances from the task with decision included, and utterances from the task without decision. This time, participants were asked to rate each utterance on 10 5-point scales (with "1" meaning "strongly disagree" and "5" meaning "strongly agree"), as to whether the utterances sounded *friendly, hesitant, concerned, smiling, certain, nervous, sad, pleasant, interested* or *monotone*. These labels were chosen based on some pilot observations. In half of the cases, the highest score on the scale represented a negative connotation (like *nervous* or *monotone*), in the other half a relatively positive one (like *smiling* or *certain*).

B. Participants

Thirty students from the humanities faculty of Tilburg University participated in this next perception experiment. They were all native speakers of Dutch, and participated as a partial fulfillment for course credit. Further demographic information was not recorded. None of them had participated in any of the previous experiments of this study.

C. Results and discussion

Table IV presents results of subject ratings of the emotional content of stimuli collected under all four conditions (positive and negative decision, with or without decision included later in message).

The table reveals some differences in perceived qualities of utterances in positive and negative mood conditions, but only for the utterances elicited in the condition in which the decision was later included. This finding is similar to results presented in Sec. IV, where reliable differences in valence were found only for this condition. According to our judges, the utterances collected in the negative valence condition (the job was not offered) sounded significantly more concerned, contained less evidence of smiling, and sounded more sad and more monotonous. There were also trends for these utterances to sound more nervous and less interested. Again in line with the results of the earlier perception test for valence, the utterances of the positive and negative conditions of the experiment without message were not judged to have significantly different perceived qualities. And again, recall that speakers in both production conditions (with or without decision in message) showed comparable moods after mood induction, whether positive or negative.

As an additional check to verify whether listeners had used those voice attributes as cues for their judgments of upcoming good or bad news, we computed correlations between the attributes and the percentage of good news responses for the stimulus utterances (see Table V). Note that these good news responses are not always "correct," as participants' judgments may be false. The table suggests that participants may have used the voice attributes in their judgments, as all of them correlated significantly (either positively or negatively) with the percentage of good news responses. More specifically, the more an utterance sounded as if it was introducing good news, the more positive the correlation with perception of the voice as friendly, smiling, certain, pleasant and interested, and the more negative the correlation with perception of the voice as hesitant, concerned, nervous, sad and monotone. Note that these effects

TABLE V. Pearson r values for correlations between the perception of good news judgments and perceived voice attributes ($^{*}p<0.05$; $^{**}p<0.01$) for stimulus utterances of the experiments with and without decision.

Decis.	Friendly	Hesitant	Concerned	Smiling	Certain	Nervous	Sad	Pleasant	Interested	Monotone
Yes	0.629**	−0.875**	−0.798**	0.778**	0.839**	−0.893**	−0.837**	0.714**	0.744**	−0.483*
No	0.706**	−0.805**	−0.613**	0.729**	0.784**	−0.703**	−0.742**	0.840**	0.664**	−0.703**

were true both for the utterances with decision and for those without decision. In the latter case, the judgments were misleading, as they did not match the actual status of the message in terms of good or bad news.

In order to find possible acoustic correlates of these voice characteristics ratings, we automatically measured a number of features from the speech signal. Using Praat scripts, we obtained some pitch-related [average, standard deviation, maximum and minimum of fundamental frequency (F_0)], and energy-related values (average, standard deviation, maximum and minimum of rms). In addition, we measured timing-related values, i.e., overall duration and speaking rate. The latter was calculated in syllables per second, where syllables were identified automatically using a Praat script written by de Jong and Wempe (2009) for detecting syllable nuclei in Dutch. This acoustic information was then correlated (using Pearson correlation) with the quality judgments presented in Table IV. Table VI presents the correlation matrix between acoustic features in each token and mean judgments of raters.

The table shows that rms-average and rms-sd were the more informative features, as they appeared to correlate with 9 of the judgments. Some of these judgments do indeed distinguish positive from negative emotions. On the other hand, F_0-related features appear to have been far less useful, although they did correlate significantly with friendly and monotone judgments.

While we did find that acoustic/prosodic features of the speech data *were* correlated strongly with raters' judgments of a number of different voice qualities, it should be noted that the same acoustic information is significantly correlated with multiple emotions. And when we examine the relationships between proportion of good news judgments and the acoustic/prosodic features of the tokens, we find fewer significant correlations, especially when we focus on the condition for which human subjects found clearest distinctions ('with message'). Table VII displays the correlations between the percentage of positive score judgments per utterance and acoustic/prosodic features. Even though only a few correlations turn out to be significant, we do see some patterns emerge from this table. For the 'with decision' condition, in which listeners reliably distinguish 'good' from 'bad' news, pitch minimum and maximum are negatively correlated with scores and speaking rate is positively correlated. That is, 'bad news' judgments tend to be uttered in a lower pitch range (lower pitch maxima and minima) and spoken more rapidly. For the 'without decision' condition, the correlations are different: pitch maximum is positively correlated; total duration is negatively correlated; intensity mean and minimum are positively correlated but maximum and standard deviation are negatively correlated; and rate is positively correlated. But these utterances do not produce the consensus we find for the 'with message' condition among our judges. So we may hypothesize that the acoustic features we find correlated with judgments for the 'with message' condition are those that are effective in conveying 'bad' vs. 'good' news.

VI. GENERAL DISCUSSION

This study has demonstrated empirical support for the impression people often have that you can tell whether a speaker is about to provide good or bad news even when there are no lexical cues present. We employed a specific elicitation procedure in which speakers were given the task of leaving a message on another person's answering machine; that message either did or did not contain information on the positive or negative decision about a job interview. Utterances collected in this way were presented to listeners, who had to decide whether a given utterance (which did not contain the actual details about the job decision) was the introduction to a positive or negative outcome of the job

TABLE VI. Pearson r values for correlations between perceived emotional content and a number of automatically obtained acoustic measures ($^{*}p<0.05$; $^{**}p<0.01$).

Judgment	Dur.	Temp.	F_0-av	F_0-max	F_0-min	F_0-sd	rms-av	rms-max	rms-min	rms-sd
Friendly	−0.229	0.320*	0.403**	0.157	0.020	0.398*	0.358*	0.039	0.321*	−0.384*
Hesitant	0.423**	−0.311	0.049	0.080	0.044	−0.023	−0.343*	0.109	−0.168	0.379*
Concerned	0.396*	−0.294	0.087	0.193	−0.026	0.012	−0.260	−0.089	0.007	0.090
Smiling	−0.265	0.264	0.203	0.034	−0.046	0.287	0.334*	0.012	0.166	−0.279
Certain	−0.368*	0.225	−0.098	−0.100	−0.033	−0.074	0.361*	−0.062	0.128	−0.366*
Nervous	0.399*	−0.310	0.111	0.131	0.044	0.087	−0.370*	0.029	−0.205	0.328*
Sad	0.470**	−0.352*	−0.001	0.078	0.148	−0.138	−0.413**	0.029	−0.230	0.342*
Pleasant	−0.232	0.288	0.205	0.017	−0.099	0.209	0.382*	−0.012	0.260	−0.375*
Interested	−0.209	0.309	0.130	0.059	−0.017	0.181	0.433**	0.116	0.335*	−0.313*
Monotone	0.258	−0.198	−0.308	−0.128	−0.020	−0.378*	−0.475**	−0.085	−0.335*	0.374*

TABLE VII. Pearson r values for correlations between the perception of good news judgments and a number of automatically obtained acoustic measures ($^*p<0.05$; $^{**}p<0.01$) for stimulus utterances of the experiments with and without decision.

Decis.	Dur.	Temp.	F_0-av.	F_0-max	F_0-min	F_0-sd	rms-av.	rms-max	rms-min	rms-sd
Yes	−0.200	0.357	−0.143	−0.347	−0.397*	0.095	0.182	−0.011	−0.052	0.024
No	−0.630**	0.052	0.038	0.355	0.113	0.062	0.467*	−0.324	0.363	−0.698**

interview. Listeners were able to perform significantly above chance level, but only for the version of an experiment in which speakers had left the actual decision on the answering machine. In the other experimental condition, in which the caller was inviting the addressee to call back without actually learning the decision from the message, listeners could not reliably determine whether the outcome would be positive or not. Additional measurements revealed that initial messages produced in 'good news' conditions emotional content could be separated from those produced in 'bad news' conditions on the basis of a number of perceptual voice attributes (e.g., whether a voice sounded as if a speaker was smiling), which in turn correlated with a number of automatically obtained acoustic/prosodic features.

An interesting outcome of the first study is that the two production tasks, leaving a message on an answering machine either with or without the actual decision of the interview, leads to quite different behavior on the part of the speaker. Whereas listeners cannot reliably derive emotional cues in utterances from the "without decision" task, they are able to do so in the "with decision" task. This difference does not seem to be related to a difference between the emotional states between speakers in the two tasks, as our mood induction procedure appeared to generate similar patterns in mood changes in both tasks (as is clear from the self-judgments). What appears to be the case instead is that other pragmatic factors determine the extent to which speakers display their mood. In this experiment, they appear only to display a perceptually positive or negative mood when they must in fact perform a positive or negative task. That is, only when they must explicitly reveal good or bad news do they display the corresponding mood in their speech.

Further research is needed on the functional relevance of prosodic predictors of emotional content. Predictive cues in general seem important for listeners, as they help to prepare the decoding of upcoming speech. We might speculate that early cues to emotional content are likely to be relevant as they guide a listener to mentally prepare for good or bad news. In addition, it would be useful to explore whether the results obtained via the experimental approach of the current study generalize to more naturalistic settings, and to what extent predictors of emotional content could be made useful in specific applications. In various kinds of interactive settings, such as in conversations with call centers (e.g., Lee and Narayanan, 2005) or with specific spoken dialog systems (Hirschberg *et al.*, 2004), it is crucial to detect certain emotional states, such as customer or user frustration, as early as possible, so that communication problems can be solved as quickly as possible. For such purposes, predictors of emotional content would obviously be very useful. In addition to the acoustic features that were shown to correlate with emotional content, it would be instructive to include other, more sophisticated ones, in particular those that are more spectral in nature, or relate to aspects of voice quality (see Truong, 2009; Bitouk *et al.*, 2010). In future work, using a larger corpus than the one employed here, it would be interesting to explore whether such measures, in combination with other acoustic features, are useful to automatically detect stretches of speech that presignal negative or positive emotional content.

ACKNOWLEDGMENTS

This research was conducted as part of a VIDI-project ("Functions of audiovisual prosody"), sponsored by the Netherlands Organisation for Scientific Research (NWO). Thanks are due to Lizette Kleij and Linda de Jong for help with the data collection, and Lennard van de Laar for assistance with setting up the perception studies.

Ang, J., Dhillon, R., Krupski, A., Shriberg, E., and Stolcke, A. (**2002**). "Prosody-based automatic detection of annoyance and frustration in human-computer dialog," in Proceedings ICSLP-2002, Denver, Colorado, pp. 2037–2040.

Bänziger, T., and Scherer, K. R. (**2005**). "The role of intonation in emotional expressions," Speech Commun. **46**, 252–267.

Barkhuysen, P., Krahmer, E. J., and Swerts, M. (**2008**). "The interplay between the auditory and visual modality for end-of-utterance detection," J. Acoust. Soc. Am. **123**, 354–365.

Batliner, A., Biersack, S., and Steidl, S. (**2006**). "The prosody of pet robot directed speech: evidence from children," Proceedings of the 3rd Speech Prosody Conference, Dresden, Germany, pp. 1–4.

Bitouk, D., Verma, R., and Nenkova, A. (**2010**). "Class-level spectral features for emotion recognition," Speech Commun. **52**, 613–625.

Brennan, S. E., and Williams, M. (**1995**). "The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers," J. Mem. Lang. **34**, 383–398.

Carlson, R., Hirschberg, J., and Swerts, M. (**2005**). "Cues to upcoming Swedish prosodic boundaries: Subjective judgment studies and acoustic correlates," Speech Commun. **46**, 326–333.

Cutler, A., Dahan, D., and van Donselaar, W. (**1997**). "Prosody in the comprehension of spoken language: A literature review," Lang Speech **40**, 141–201.

Dahan, D., Tanenhaus, M. K., and Chambers, C. G. (**2002**). "Accent and reference resolution in spoken-language comprehension," J. Mem. Lang. **47**, 292–314.

de Jong, N., and Wempe, T. (**2009**). "Praat script to detect syllable nuclei and measure speech rate automatically," Behav. Res. Methods Instrum. Comput. **41**, 385–390.

Ekman, P., and Friesen, W. V. (**1975**). *Unmasking the Face: A Guide to Recognizing Emotions from Facial Clues* (Prentice Hall, New Jersey).

Fox Tree, J. E. (**2001**). "Listeners uses of um and uh in speech comprehension," Mem. Cognit. **29**, 320–326.

Gravano, A., and Hirschberg, J. (**2009**). "Turn-Yielding Cues in Task-Oriented Dialogue," in Proceedings of the SIGDIAL2009 Conference, London, UK, pp. 253–261.

Grosjean, F. (**1983**). "How long is the sentence? Prediction and prosody in the on-line processing of language," Linguistics **21**, 501–530.

Hart, J. T., Collier, R., and Cohen, A. (**1990**). *A Perceptual Study of Intonation* (Cambridge University Press, Cambridge).

Hirschberg, J., Litman, D., and Swerts, M. (**2004**). "Prosodic and other cues

to speech recognition failures," Speech Commun. **43**, 155–175.
Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A., and Den, Y. (**1998**). "An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese map task dialogs," Lang Speech **41**, 295–321.
Krahmer, E. J., and Swerts, M. (**2001**). "On the alleged existence of contrastive accents," Speech Commun. **34**, 391–405.
Krahmer, E. J., van Dorst, J., and Ummelen, N. (**2004**). "Mood, persuasion and information presentation: The influence of mood on the effectiveness of persuasive digital documents," Inf. Des. J. **12**, 40–52.
Ladd, D. R. (**1996**). *Intonational Phonology* (Cambridge University Press, Cambridge).
Lee, C. M., and Narayanan, S. S. (**2005**). "Toward detecting emotions in spoken dialogs," IEEE Trans. Speech Audio Process. **13**, 293–303.
Leroy, L. (**1984**). "The psychological reality of fundamental frequency declination," *Antwerp Papers in Linguistics* (UA University Press, Antwerp), 102 pages.
Levelt, P. M. (**1989**). *Speaking: From Intention to Articulation* (MIT Press, Cambridge, MA).
Levinson, S. (**1983**). *Pragmatics* (Cambridge University Press, Cambridge).
Liscombe, J., Hirschberg, J., and Venditti, J. J. (**2005**). "Detecting certainness in spoken tutorial dialogues," in Proceedings of the International Conference on Spoken Language Processing (Interspeech 2005), Lisbon, Portugal, pp. 1837–1840.
Mozziconacci, S. J. L. (**1998**). "Speech variability and emotion: Production and perception," Ph.D. thesis, Technical University Eindhoven, The Netherlands.
Pon-Barry, H. (**2008**). "Prosodic manifestations of confidence and uncertainty in spoken language," in Proceedings of the International Conference on Spoken Language Processing (Interspeech 2008), Brisbane, Australia, pp. 74–77.
Pon-Barry, H., and Shieber, S. (**2009**). "Identifying uncertain words within an utterance via prosodic features," in Proceedings of the International Conference on Spoken Language Processing (Interspeech 2009), Brighton, UK, pp. 1579–1582.
Scherer, K. R. (**2003**). "Vocal communication of emotion: A review of research paradigms," Speech Commun. **40**, 227–256.
Schuller, B., Vlasenko, B., Eyben, F., Rigoll, G., and Wendemuth, A. (**2009**). "Acoustic emotion recognition: A benchmark comparison of performances," in Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Merano, Italy, pp. 552–557.
Silverman, S., and Pierrehumbert, J. (**1990**). "The timing of prenuclear high accents in English," in *Laboratory Phonology, Vol I: Between the Grammar and Physics of Speech*, edited by J. Kingston and M. Beckman (Cambridge University Press, Cambridge), pp. 71–106.
Smith, V. L., and Clark, H. H. (**1993**). "On the course of answering questions," J. Mem. Lang. **32**, 25–38.
Swerts, M. (**1998**). "Filled pauses as markers of discourse structure," J. Pragmat. **30**, 485–496.
Swerts, M. (**2007**). "Contrast and accent in Dutch and Romanian," J. Phonetics **35**, 380–397.
Swerts, M., Collier, R., and Terken, J. (**1994**). "Prosodic predictors of discourse finality in spontaneous monologues," Speech Commun. **15**, 79–90.
Swerts, M., and Krahmer, E. J. (**2005**). "Audiovisual prosody and feeling of knowing," J. Mem. Lang. **53**, 81–94.
Swerts, M., and Krahmer, E. J. (**2010**). "Visual prosody of newsreaders: Effects of information structure, emotional content and intended audience," J. Phonetics **38**, 197–206.
Truong, K. (**2009**). "How does react affect affect affect recognition in speech?," Ph.D. thesis, University of Twente, The Netherlands.
Umeda, N. (**1982**). "F_0 declination is situation dependent," J. Phonetics **10**, 279–290.
Ward, N., and Tsukahara, W. (**2000**). "Prosodic features which cue back-channel responses in English and Japanese," J. Pragmatics **32**, 1177–1207.
Wilting, J., Krahmer, E. J., and Swerts, M. (**2006**). "Real vs. acted emotional speech," in Proceedings of the International Conference on Spoken Language Processing (Interspeech 2006), Pittsburgh, PA, pp. 805–808.
Yuan, J., Shen, L., and Chen, F. (**2002**). "The acoustic realization of anger, fear, joy and sadness in Chinese," in Proceedings ICSLP-2002, Denver, CO, pp. 2025–2028.