Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 25 (2011) 84–104

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Expression of affect in spontaneous speech: Acoustic correlates and automatic detection of irritation and resignation

Petri Laukka [a,b,*], Daniel Neiberg [c], Mimmi Forsell [c], Inger Karlsson [c], Kjell Elenius [c]

[a] *Department of Psychology, Uppsala University, Uppsala, Sweden*
[b] *Department of Education and Psychology, University of Gävle, Gävle, Sweden*
[c] *Centre for Speech Technology, Department of Speech, Music and Hearing, KTH, Stockholm, Sweden*

Received 30 May 2008; received in revised form 21 January 2010; accepted 23 March 2010
Available online 1 April 2010

Abstract

The majority of previous studies on vocal expression have been conducted on posed expressions. In contrast, we utilized a large corpus of authentic affective speech recorded from real-life voice controlled telephone services. Listeners rated a selection of 200 utterances from this corpus with regard to level of perceived irritation, resignation, neutrality, and emotion intensity. The selected utterances came from 64 different speakers who each provided both neutral and affective stimuli. All utterances were further automatically analyzed regarding a comprehensive set of acoustic measures related to F0, intensity, formants, voice source, and temporal characteristics of speech. Results first showed that several significant acoustic differences were found between utterances classified as neutral and utterances classified as irritated or resigned using a within-persons design. Second, listeners' ratings on each scale were associated with several acoustic measures. In general the acoustic correlates of irritation, resignation, and emotion intensity were similar to previous findings obtained with posed expressions, though the effect sizes were smaller for the authentic expressions. Third, automatic classification (using LDA classifiers both with and without speaker adaptation) of irritation, resignation, and neutral performed at a level comparable to human performance, though human listeners and machines did not necessarily classify individual utterances similarly. Fourth, clearly perceived exemplars of irritation and resignation were rare in our corpus. These findings were discussed in relation to future research.

Keywords: Acoustic features; Automatic speech classification; Emotion recognition; Human–computer interaction; Spontaneous speech

1. Introduction

Whenever people communicate with each other through speech, they always convey affective expression non-verbally through their tone of voice (i.e., vocal expression) apart from the basic meaning conveyed by

[*] Corresponding author. Address: Department of Psychology, Stockholm University, 106 91 Stockholm, Sweden. Tel.: +46 8 163935; fax: +46 8 159342.
E-mail address: petri.laukka@psychology.su.se (P. Laukka).

0885-2308/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2010.03.004

the actual words they use. Vocal expression of affect has received much attention from speech scientists in recent years and has important implications for speech and language research and technology (e.g., Tatham and Morton, 2004). For example, knowledge of how affect is conveyed in speech is important for the improvement of automatic speech recognition – both for improving the recognition of the linguistic content of affective speech (Athanaselis et al., 2005) and the recognition of the speaker's affective state (Devillers et al., 2005; Shami and Verhelst, 2007). Also, adding vocal expression to synthetic speech would improve the naturalness of the speech produced (Schröder, 2001; Murray and Arnott, 2008). Therefore research on vocal expression has implications for many aspects of affective computing, like developing human machine interfaces that are adaptive and can respond to a user's behavior. Seen from a broader perspective, the communication of affect, including vocal expression, is also of major importance for the regulation of human social behavior and interaction (Buck, 1984; Darwin, 1872/1998; Ekman, 2003; Russell et al., 2003).

Several reviews of studies of vocal expression have been published in recent years (e.g., Cowie et al., 2001; Juslin and Laukka, 2003; Laukka, 2008; Scherer, 2003). The results of these reviews converge on the finding that specific emotions like anger, fear, joy, and sadness can be accurately communicated non-verbally through the voice. Also, the acoustical correlates of each emotion seem to be relatively distinct (e.g., Juslin and Laukka, 2003, Tables 7 and 8). However, there are a number of limitations with previous research. First, the majority of previous studies on vocal expressions – and emotion expression in general – have been conducted on posed expressions. In a typical study, actors have been asked to portray expressions and listeners have then been asked to identify the portrayed expressions in recognition tests. Research on posed expressions has produced many important findings, and it can reasonably be argued that posed expressions must be relatively similar to naturally occurring expressions in order for communication to be successful (e.g., Davitz, 1964). Nevertheless posed expressions may also be exaggerated and more intense and prototypical than authentic everyday expressions (Scherer, 1986; see also Laukka et al., 2009). There also exists the possibility that some aspects of the voice which convey affective information are not under voluntary control (e.g., Bachorowski and Owren, 1995). Second, previous research has mostly studied the expression of prototypical expressions of full-blown emotions (e.g., anger, fear, joy, sadness), defined as "relatively brief episodes of synchronized responses of all or most organismic subsystems in response to the evaluation of an external or internal event as being of major significance" (Scherer, 2003, p. 243, Table 4). Emotions per se are thus regarded as intense affective states. In contrast, studies on spontaneous affect expression in everyday speech have reported that prototypical expressions of full-blown emotions are rarely found in normal day-to-day conversations; whereas expressions of milder and more subtle affective states are more frequently occurring (Campbell, 2005; Cowie, 2009; Cowie and Cornelius, 2003; Devillers et al., 2005). Thus, it is important to conduct further studies on authentic expressions of affective states other than full-blown emotions. Hence, in the present study, we investigated the acoustic correlates and automatic detection of two relatively mild affective states – irritation and resignation – from authentic affective speech collected from real life human–computer interactions.

2. Previous research on authentic affective speech

Researchers have attempted to study authentic vocal expressions in various ways. For example, some researchers have used various affect induction methods in order to study the effects of the manipulation on the voice (e.g., Aubergé et al., 2006; Bachorowski and Owren, 1995; Barrett and Paus, 2002; Bonner, 1943; Johnstone et al., 2005, 2007; Laukka et al., 2008). These kinds of studies are valuable, but also have limitations. For one thing, it is difficult to induce strong and well-differentiated emotional reactions in laboratory settings, which makes the study of intense emotional reactions difficult. The resulting speech may also be more or less similar to everyday conversations depending on which affect induction method is used. For example, using traditional mood induction methods in laboratory settings may lead to artificial speech, whereas methods where the speakers interact with humans or computers may lead to less restrained speech (Batliner et al., 2003a, 2008). Another route for studying authentic expressions is therefore to investigate recordings of spontaneous speech from various real-life affective situations (e.g., Devillers et al., 2005; Douglas-Cowie et al., 2003; Eldred and Price, 1958; Greasley et al., 2000; Lee and Narayanan, 2005; Litman and Forbes-Riley, 2006; Williams and Stevens, 1972). However, the study of real-life conversations is often marred by a lack of control over what affective states – if any – the speakers actually were experiencing during the recorded situations. It is also difficult to obtain recordings of naturally occurring voice changes in affective situations for large enough numbers of speakers to allow for statistical analyses of the material.

Another complication for the study of authentic expressions is that any naturally occurring expression is shaped to a certain degree by both the somatic alterations caused by the emotional reaction (i.e., *push effects*) as well as external conditions such as social norms and cultural display rules (i.e., *pull effects*; Ekman and Friesen, 1969; Scherer, 1989). The combination of both push and pull effects may lead to strategically posed expressions or to the masking of expressions, which may contribute to the subtlety of authentic expressions.

3. Prior studies on automatic recognition of affect from vocal expression

The automatic recognition of speakers' affective states based on acoustic voice analyses is a concern of great interest for affective computing. Most previous studies have been conducted on posed expressions, and have used a variety of recognition systems with varying degrees of automaticity and based on different types of machine learning architectures or statistical methods (e.g., Banse and Scherer, 1996; Dellaert et al., 1996; Nicholson et al., 2000; Nogueiras et al., 2001; Nwe et al., 2003; Oudeyer, 2003; Petrushin, 1999; Toivanen et al., 2004). Such studies have shown that it is possible to predict the speakers' intended emotions (including, but not limited to, anger, fear, happiness, and sadness) from the acoustic characteristics of the portrayed expressions with accuracy better than chance. Automatic discrimination of affective states is usually less successful than human discrimination performance (see Ververidis and Kotropoulos, 2006, for a review), but some recent studies employing a wide variety of acoustic features (including temporal modeling) have achieved very high recognition rates (e.g., Kandali et al., 2009; Schuller and Rigoll, 2006).

A number of studies have also attempted to use automatic detection systems for the classification of speaker states from real-life corpora. Such corpora may consist of speech collected from, for example, spoken tutoring dialogues (Litman and Forbes-Riley, 2006), interactions with multi-modal dialogue systems (Vogt and André, 2005), television talk shows (Grimm et al., 2007), call-centers (Burkhardt et al., 2008; Morrison et al., 2007), or robot–human interaction (Batliner et al., 2008). Compared with studies of posed expressions, studies on automatic detection of authentic expressions generally include fewer categories, because each speech corpus often contains limited varieties of affective speech. Most commonly studies have sought to classify speech into categories of positive and negative affect (e.g., Chateau et al., 2004; Lee and Narayanan, 2005; Litman and Forbes-Riley, 2006), or angry/frustrated and neutral (Ang et al., 2002; Burkhardt et al., 2006, 2008; Morrison et al., 2007), though Vidrascu and Devillers (2007) reported automatic detection of five affective categories (fear, anger, sadness, neutral, and relief). A couple of studies have also studied the automatic detection of affective states such as interest (Schuller et al., 2007), approval, attention and prohibition (Breazeal and Aryananda, 2002; Slaney and McRoberts, 2003), and motherese and emphatic speech (Batliner et al., 2006). Most studies report better than chance detection, but generally, the closer one gets to natural speech, the less clear the emotion expressions become and the harder the task of automatically classifying them becomes (Batliner et al., 2003a; Vogt and André, 2005).

4. Contribution of the present study

In the present study, we investigated the acoustic correlates and automatic detection of affect using authentic affective speech recorded from voice controlled telephone services – thereby complementing earlier research which has mainly been conducted on posed expressions. More specifically, we focused on human–computer interactions where the communication was not working, and have targeted episodes where the speakers become either irritated or resigned because the computer interface does not understand the speakers' instructions. Thus, the context of our recordings allowed a rather high level of control over the speakers' probable affective states. We also only included speakers who contributed both neutral and affective utterances. This way we were able to control for individual differences in baselines between acoustic measures, whereas a lot of previous research has failed to include adequate control conditions with which to compare affective speech.

Irritation and resignation belong to the same broad emotion families as the more commonly investigated full-blown emotions anger and sadness, respectively (e.g., Ekman, 1992). However, the labels anger and sadness are generally reserved for very intense reactions to certain prototypical situations, often called basic or

Table 1
Brief summary of selected acoustic characteristics of anger and sadness from previous studies obtained with (mainly) posed expressions (based on Juslin and Laukka, 2003).

Acoustic measure	Anger	Sadness
F0 mean	+	–
F0 variability	+	–
Intensity mean	+	–
Intensity variability	+	–
F1 mean	+	–
F1 bandwidth	–	+
High-frequency energy	+	–
Speech rate	+	–

Note: + = high; – = low.

modal emotions, whereas the affective states in our corpus were much less intense. Within these broad emotion families, irritation and anger – and resignation and sadness – also share the same patterns of emotion eliciting cognitive appraisals (Scherer and Tran, 2001). Therefore, we also compared the acoustic correlates of authentic expressions of irritation and resignation with the acoustic correlates of anger and sadness obtained from (a) earlier studies conducted on mainly posed expressions and (b) Scherer's (1986) theoretical predictions based on somatic alterations associated with emotional responses (see Table 1 for a summary of these acoustic correlates, based on the review by Juslin and Laukka, 2003).[1]

All utterances were validated in listening tests and subjected to detailed automatic acoustic analyses to allow for the study of the acoustical correlates of authentic affective speech. The listening tests assessed both: (a) what affective state each utterance was perceived as expressive of (i.e., irritation, resignation, or neutral) and (b) how intense the perceived affective state of each utterance was. Previous research has shown that listeners can accurately recognize not only broad emotion categories, but also finer nuances within these categories from vocal expressions. Banse and Scherer (1996) showed that listeners could discriminate between different members of the same emotion family (e.g., irritation and rage, see also Frick, 1986). Recent studies have further shown that listeners can accurately recognize the intensity of the emotion expressed (e.g., Juslin and Laukka, 2001; Laukka and Juslin, 2007). The relative intensity of affective states is of great importance for the behavioral and physiological emotional responses (Brehm, 1999; Sonnemans and Frijda, 1994), and can thus have a great impact on the acoustic characteristics of various expressions. For example, Juslin and Laukka (2001) showed that in some cases differences in acoustic measures were actually larger between different intensities of the same emotion category than between different emotion categories. However, no previous study has investigated the perception and acoustical correlates of emotion intensity from authentic vocal expressions.

When conducting studies on automatic detection, there is always a tradeoff between having a carefully validated corpus of a limited size, or to use a larger corpus annotated by only one or a few listeners. Emotion recognition ability is a skill that shows large individual differences due to, for example, gender (Hall, 2006), age (Laukka and Juslin, 2007), and probably also personality characteristics (though much work remains to be done on this particular subject). Therefore it is often the case in previous research that the more robust the classifiers are (because they have been trained on a large corpus), the more uncertain it is what is really being classified (because the corpus has only been coded by a few coders). In the present study, we therefore

[1] Based on his component process theory, Scherer (1986) made detailed predictions about the patterns of acoustic measures associated with different emotions. The predictions were based on the idea that emotions involve sequential cognitive appraisal, or stimulus evaluation checks, of stimulus features such as novelty, intrinsic pleasantness, goal significance, coping potential, and norm/self compatibility. The outcome of each specific stimulus evaluation check is assumed to have an effect on the physiological responding in ways that influence voice production. For example, anger yields increased tension in the laryngeal musculature coupled with increased subglottal air pressure, which will change the production of sound at the glottis and hence change the timbre of the voice. Scherer's predictions have received some empirical support from studies of posed expressions (e.g., Banse and Scherer, 1996; Juslin and Laukka, 2001). The results in Table 1 are in accordance with Scherer's predictions for hot anger and sadness, respectively (except for the finding that sadness is most often associated with low mean F1, whereas high mean F1 was predicted). Though the coupling between physiological emotional responding and resulting changes in the voice is in theory well established, there is yet relatively little empirical evidence from studies of authentic affective speech to support it (e.g., Johnstone et al., 2007).

used a larger than usual number of listeners in order to get reliable recognition data for our corpus. It should be noted that it is especially important to have carefully validated data in studies of spontaneous speech, because the listeners' perceptions often constitute the only available criteria for classifying stimuli.

Finally, studies on the acoustical correlates of vocal expressions, and studies on automatic detection of speaker affect, have mostly been conducted in separate research traditions and have been driven by different motivations. On the one hand, studies on automatic detection often have the applied motivation of advancing affective computing, and therefore the focus of these studies seldom lies in reporting detailed data on the acoustic correlates of affective states. On the other hand, many studies that focus on the acoustical correlates from the perspective of advancing emotion theory would benefit from the machine learning based methods of handling large amounts of acoustical data used in studies of automatic detection. In the present study, we pursued both of the above strands of research, and also compared human perception of irritation and resignation with automatic classification performance. Because each speaker provided both neutral and affective utterances, we also compared classifiers with, and without, speaker adaptation.

5. Method

5.1. Speech material

The speech material used was telephone speech recorded at 8 kHz from real life human–computer interactions during November 2005 by the Swedish company *Voice Provider* that runs various voice controlled telephone services covering, for example, information regarding airlines, ferry traffic, and postal services. The original database consisted of 61,078 utterances of varying length; mainly brief commands, such as "ja" [yes], "nej" [no], and other short words in Swedish (e.g., "godkänd" [approved], "kundservice" [customer service], "privatperson" [private customer], "företag" [company], "biljetter" [tickets], and various Swedish town names). All utterances were originally classified into emotionally neutral, emotionally negative or emphatic speech by a senior speech researcher (I.K.). This first classification revealed that 96.1% of the utterances did not contain affective speech and were classified as neutral, whereas 1.7% were classified as emphatic and 2.2% were classified as negative. Utterances classified as emphatic were pronounced in an emphatic, or hyper-articulated, speech-style without necessarily being emotional (e.g., Selting, 1994), and the utterances classified as negatively valenced mainly included irritation and resignation. Parts of this corpus have been used in prior studies on the automatic detection of affect from speech (e.g., Neiberg et al., 2006; Neiberg and Elenius, 2008).

For our purposes we made a selection of utterances from the Voice Provider database with the constraint that we should have at least one neutral and one affective utterance from each included speaker, in order to allow for within-subjects analyses. A further constraint was that the utterances should be of sufficient recording quality for acoustic analysis and not contain any truncations, repetitions or other problems. Finally, a few utterances that included linguistic emotional content (e.g., swear words) were excluded, which left us with 261 utterances from which we made the final selection of 200 stimuli for inclusion in the listening tests (to keep the time required for the listening tests reasonable). The selected 200 utterances came from 64 different speakers and had the following original classifications: 81 neutral, 31 emphatic, 21 resigned and 67 irritated stimuli. Each speaker contributed between 2 and 6 utterances. Preliminary analyses of the acoustical characteristics of this selection have been reported in Forsell (2007).

5.2. Acoustic measures

Seventy-three acoustic measures related to pitch, intensity, formants, voice source and duration were automatically extracted using Praat scripts (Boersma and Weenink, 2007), see Table 2 for a brief description of each acoustic measure.[2] In a preprocessing step, each utterance was segmented into pseudo-syllables, defined

[2] Note that we did not use inverse filtering for the measurement of voice source measures, but instead used the following spectral approximations: *open quotient* = F0 amplitude-2nd F0 harmonic amplitude; *glottal opening* = F0 amplitude-F1 amplitude; *amplitude of voicing* = F0 amplitude; *rate of closure* = F0 amplitude-F3 amplitude; *skewness* = F0 amplitude-F2 amplitude; *completeness of closure* = F1 mean bandwidth (Touboul, 2003).

Table 2
Acoustic measures related to pitch, intensity, formants, voice source and temporal aspects of speech.

Abbreviation	Comment
Pitch cues	
F0Min/Max/M/Range/SD	Minimum, maximum, mean, range, standard deviation of F0
F0MinRelPos, F0MaxRelPos	Relative position of the minimum and the maximum of F0
F0Q1-5	Quantiles 1–5 of F0
F0Slope	Slope of F0
F0DeltaM, F0DeltaDeltaM	Frame-wise delta, frame-wise delta–delta of F0
F0FracRise/Fall/Stat	Percentage of frames with F0 rise/fall/ stationary
F0SSubtMin/Max/Range/SD	Minimum, maximum, mean, range, standard deviation of F0 with the slope subtracted
F0SSubtQ1–5	Quantiles 1–5 of F0 with the slope subtracted
FracVoiced	Percentage of voiced frames
F0AtIntMax	F0 at intensity maximum
Jitter	The average absolute difference between consecutive differences for consecutive periods, divided by the average period
Intensity cues	
IntMax/M/Range/SD	Minimum, maximum, mean, range, standard deviation of Intensity
IntMaxRelPos	Relative position of the maximum of F0
IntQ1–5	Quantiles 1–5 of intensity
IntSlope	Slope of intensity
IntFracRise/Fall/Stat	Percentage of frames with intensity rise/fall/ stationary
IntDeltaM, IntDeltaDeltaM	Frame-wise delta, frame-wise delta–delta of intensity
IntAtF0Max/Min	Intensity at F0 maximum and minimum
Shimmer	The average absolute difference between consecutive differences for amplitudes of consecutive periods
Formant cues	
F1–4M	Mean of formants 1–4
F1–4SD	Standard deviation of formants 1–4
F1–4B	Median bandwidth of formants 1–4
Voice source cues	
H1MH2	F0 amplitude-2nd F0 harmonic amplitude
H1MA1–3	F0 amplitude-formant 1–3 amplitude
H1A	F0 amplitude
Temporal cues	
SilenceDurM	Mean of silence duration
SyllableDurM	Mean of syllable duration

as sequences of unvoiced, voiced, and unvoiced segments with intensity and duration above certain thresholds. Most acoustic measures were calculated for these pseudo-syllables, and their averages over the whole utterance were then used in our statistic analyses. Some fundamental frequency and intensity based measurements, such as minimum, maximum, range, mean, standard deviation, quantiles and slope were measured over the whole utterance. All pitch based measures use logarithmic scale and all intensity based measures use dB scale.

However, including all 73 acoustic measures in statistical analyses would lead to a risk for alpha-inflation. We therefore used principal components analysis (PCA) with oblimin rotation in order to reduce the number of variables to be included in the analyses. We utilized oblimin rotation because previous research has showed that the acoustical measures are often correlated in emotional corpora (e.g., Banse and Scherer, 1996; Juslin and Laukka, 2001). Multiple criteria were used to decide on the appropriate number of factors to retain: scree test, the latent root criterion (e.g., eigenvalues of 1 or greater), and the interpretability of solutions (see Hair et al., 1998; Zwick and Velicer, 1986). The highest loading measures for each factor were chosen for further analyses (i.e., F0M, F0SD, F0FracFall, IntM, IntSD, IntFracRise, H1MH2, F2M, F1B, IntFracFall, F3M, Shimmer, SyllableDurM). After studying the intercorrelations between the acoustic measures for each factor, we included some further cues in order to better represent each factor (F0Q1, F0Q5, F0FracRise, IntQ1, IntQ5, H1MA3, F1M, F2B, F3B, Jitter), leading to a set of 23 acoustic measures to be included in the further analyses.

5.3. Listening experiment

Twenty listeners (7 women, 13 men, mean age = 29.5 years) rated all speech samples ($N = 200$) on the following scales: irritation/anger, resignation/sadness, and neutral. All scales ranged from 0 (not perceived at all) to 7 (very clearly perceived). The annotation of broad emotion classes/families is a common procedure in studies on real-life expressions (e.g., Scherer and Ceschi, 2000). The listeners were also asked to rate the emotion intensity of each speech sample on a scale from 0 (very weak intensity) to 7 (very strong intensity). The listening tests were conducted individually using custom software, and the participants listened to the stimuli through high-quality headphones. The presentation order of the stimuli was randomized for each listener. Further, there were no time constraints and the listeners were allowed to listen to each stimulus as many times as necessary to make their judgments.

The listeners were given the following instructions: “You shall first rate each phrase on three scales describing different affective states: irritation/anger, resignation/sadness, and neutral. If a phrase, in your opinion, expresses a certain affective state very clearly you should give that phrase a rating of 7 on the corresponding scale. If a phrase, in your opinion, does not express a certain affective state at all, you should give that phrase a rating of 0 on the corresponding scale. If you think that a phrase expresses several affective states at once, you should rate it according to how clearly you perceive the different affects. ‘Neutral’ refers to an expression that is emotionally neutral and does not express any affective state. Emotions and affective states can vary with regard to how intense they are; for example, it is possible to be just a little happy or to be very happy. Therefore you should also rate how intense the affect expressed by each phrase is. If, in your opinion, the affective state which is expressed by a phrase is very strong and intense, you should give it a rating of 7 on the emotion intensity scale. If you instead think that the phrase expresses an affective state with very weak intensity, you should give it a rating of 0 on the emotion intensity scale. Note that it is possible for a phrase to express an affective state that is not very clearly perceived (i.e., it receives low ratings of irritation, resignation or neutral), but still receive a high rating on emotion intensity. Similarly it is possible for a phrase to express a certain affective state very clearly, but with weak emotion intensity”.

5.4. Automatic detection of affect

We aimed to investigate how well the affective content of the utterances in our corpus could be detected from the automatically extracted acoustic measures, and utilized the results from the listening test described above as criteria for determining which utterances were to be categorized as irritated, resigned or neutral. It seems valid to assume that acoustic features derived from this corpus are not only speaker dependent, but also dependent on channel transmission and environment. Therefore, we wanted to compare automatic detection with and without speaker adaptation. For this purpose all speakers, for whom no utterances were perceived as neutral in the listening test, were removed and one neutral utterance per speaker was used for adaptation by mean subtraction. Because the absolute numbers of utterances falling into each expressive class were quite low we chose to use linear discriminant analysis (LDA) (Fukunaga, 1990) which is a robust classifier.[3] Also, we opted for the use of the reduced set of 23 acoustic measures for the automatic detection because the number of resigned utterances was rather low, and it also facilitated the speaker adapted analyses.

[3] Our motivation for using linear discriminant analysis (LDA) as classifier was: (1) it is discriminative and robust to low number of data samples. (2) All model parameters can be directly estimated from data, i.e., there are no extra parameters such as regularization parameters or model topology that need to be optimized or given a priori. (3) In the LDA framework it is possible to specify a prior distribution of classes. For example, the distribution may be empiric or uniform. For our experiments, we wanted to assume a uniform prior distribution simply because the prior distribution is dependent on the scenario/situation, and unless a uniform distribution is assumed the results cannot be compared against evaluations done on other corpora.

Table 3
Summary of the results from the listening test.

	Classification results		
	Irritation	Resignation	Neutral
Mean rating (M/SD)			
Irritation	4.85 (0.79)	2.48 (0.96)	1.99 (0.78)
Resignation	1.78 (0.69)	4.53 (0.71)	1.68 (0.70)
Neutral	2.51 (0.68)	2.67 (0.52)	4.57 (0.67)
Emotion intensity	4.95 (0.79)	3.75 (0.66)	3.19 (0.55)
N speech samples	36	23	133
N speakers	31	15	61

For detection, we used 3-fold cross-validation. In short, the corpus was split into three sets with roughly equal distribution of affective classes. Then two sets were used for development and one for evaluation in a combinatorial procedure. The prior probabilities of the affective categories were set to a uniform distribution. All 23 acoustic measures could not be used for modeling because the problem would then be ill-posed. Therefore, a brute force forward selection procedure was adopted to find the most discriminative acoustic measures. Essentially, each acoustic measure was evaluated on the development set and the one that maximized average recall (the mean of the diagonal of the confusion matrix) was kept in an incremental way. This resulted in one list of ranked acoustic measures per cross-validation cycle. The rank of each acoustic measure was determined as the relative position, measured from the bottom to the top in the list produced by the forward selection scheme, and the final ranking was calculated as the mean for all cross-validation tests.

6. Results

6.1. Listening test

To begin with, we wanted to make sure that the 20 listeners were able to rate the speech stimuli in a consistent fashion. Therefore we calculated the *average measure intraclass correlation* for each rating scale, using the Spearman–Brown formula.[4] The inter-rater reliability was high for all scales, as evidenced by the following average measure intraclass correlations: irritation ($R = 0.93$), resignation ($R = 0.92$), neutral ($R = 0.88$), and emotion intensity ($R = 0.87$).

The results from the listening test are shown in Table 3. We used majority voting to decide which stimuli were classified as irritated, resigned, or neutral. A speech stimulus was classified as, for example, irritated if it was rated higher on the irritation scale than on the resignation or neutral scales by a majority of the listeners. Majority voting is a robust method because it is insensitive to individual differences in the baseline of ratings, and to any possible inability of the listeners to use an interval scale for the rating. Of the 200 speech stimuli, 133 were classified as neutral, 36 as irritated and 23 as resigned. Additionally, five stimuli were classified as a mix between irritation and neutral (i.e., equally many listeners gave the highest ratings to irritation and neutral), and three stimuli were classified as a mix between resignation and neutral. No utterances were rated as a mix between irritation and resignation. The numbers of speakers who provided stimuli for each affect category are also shown in Table 3.

The listeners' mean ratings on each scale are also shown in Table 3 as a function of classified affect. As intended, the classified emotion received significantly higher ratings on the corresponding scale than on the other scales (t-tests, p's $< .0001$). Regarding emotion intensity, the utterances classified as irritated received significantly higher intensity ratings than resigned utterances, which in turn received significantly higher ratings than the neutral utterances (t-tests, p's $< .0001$). The intercorrelations among the rating scales were as follows: irritation/resignation ($r = -.03$, ns), irritation/neutral ($r = -.78$, $p < .0001$), irritation/emotion intensity

[4] The average measure intraclass correlation (R) is calculated using the following formula, where r is the average inter-rater correlation across all stimuli, and n is the number of raters: $R = (r*n)/(1 + ((n - 1)*r))$.

($r = .82$, $p < .0001$), resignation/neutral ($r = -.50$, $p < .0001$), resignation/emotion intensity ($r = .01$, ns), neutral/emotion intensity ($r = -.74$, $p < .0001$). These correlations indicate that ratings of irritation and emotion intensity were associated, and that high ratings of irritation and neutral, or neutral and emotion intensity, seldom occurred together.

6.2. Acoustical correlates of authentic expressions

6.2.1. Acoustic differences between affective and neutral utterances

One of the main research questions addressed in this study was to investigate possible acoustic differences between authentic affective utterances and neutral utterances. For this purpose, we first calculated the mean values of the 23 selected acoustic measures for each speaker across all speech samples that were rated as irritated, resigned or neutral by the listeners. Each mean value was based on 1–4 utterances for each speaker. This allowed us to compare the acoustic characteristics of neutral, irritated, and resigned speech within persons, thereby controlling for individual differences in baselines of the acoustic parameters between different speakers. Then we investigated whether the mean differences between neutral and irritated/resigned utterances were significant using within-groups *t*-tests. Outliers were detected using Grubb's test and removed prior to the analyses.

The results of the comparisons are shown in Table 4 (for the comparison between speech perceived as irritated and speech perceived as neutral) and Table 5 (for the comparison between speech perceived as resigned and speech perceived as neutral). Regarding acoustic measures related to pitch, irritated speech was found to have higher F0M, F0Q1 (first quantile of F0; approximately minimum F0) and F0Q5 (fifth quantile of F0; approximately maximum F0) than neutral speech, whereas resigned speech had lower F0M and F0Q5 than neutral speech. Resigned utterances also had smaller F0SD than neutral utterances. For the measures related to intensity, irritated speech had higher IntM and IntQ5 (fifth quantile of intensity; approximately maximum

Table 4
Results of the comparisons of the acoustic characteristics of speech perceived as irritated and speech perceived as neutral.

Acoustic measure	Irritated (M/SD)	Neutral (M/SD)	t
F0M	2.29 (0.10)	2.22 (0.10)	4.88***
F0Q1	2.18 (0.11)	2.12 (0.11)	2.88**
F0Q5	2.37 (0.11)	2.31 (0.12)	4.90***
F0SD	0.074 (0.033)	0.079 (0.036)	−0.51 ns
F0FracRise	0.335 (0.148)	0.317 (0.134)	0.62 ns
F0FracFall	0.397 (0.119)	0.424 (0.132)	−0.97 ns
Jitter	0.027 (0.015)	0.025 (0.011)	0.57 ns
IntM	71.77 (6.21)	67.46 (5.33)	4.94***
IntQ1	56.13 (8.33)	53.86 (6.66)	1.34 ns
IntQ5	82.73 (5.10)	78.01 (5.05)	6.33***
IntSD	10.69 (3.10)	9.62 (2.13)	1.71 ns
IntFracRise	0.171 (0.055)	0.210 (0.065)	−2.65*
IntFracFall	0.128 (0.066)	0.117 (0.055)	0.76 ns
Shimmer	0.102 (0.039)	0.110 (0.04)	−0.79 ns
F1M	587.6 (92.9)	576.0 (71.5)	0.86 ns
F2M	1484.9 (141.1)	1479.8 (178.0)	0.16 ns
F3M	2479.7 (129.5)	2490.0 (140.04)	−0.43 ns
F1B	155.2 (76.1)	163.7 (67.8)	−0.63 ns
F2B	366.9 (180.5)	298.3 (144.2)	2.29*
F3B	381.5 (117.2)	374.7 (109.3)	0.33 ns
H1MH2	−6.77 (9.45)	−8.59 (9.78)	1.79 ns
H1MA3	8.79 (9.82)	9.52 (12.24)	−0.72 ns
SyllableDurM	0.451 (0.120)	0.412 (0.143)	2.37*

Note: $N = 24–26$.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

Table 5
Results of the comparisons of the acoustic characteristics of speech perceived as resigned and speech perceived as neutral.

Acoustic measure	Resigned (M/SD)	Neutral (M/SD)	t
F0M	2.20 (0.12)	2.26 (0.11)	−2.24*
F0Q1	2.14 (0.12)	2.19 (0.10)	−1.64 ns
F0Q5	2.27 (0.14)	2.35 (0.14)	−2.31*
F0SD	0.045 (0.022)	0.057 (0.027)	−2.15*
F0FracRise	0.208 (0.122)	0.258 (0.219)	−1.01 ns
F0FracFall	0.446 (0.185)	0.481 (0.233)	−0.66 ns
Jitter	0.031 (0.021)	0.027 (0.017)	1.02 ns
IntM	60.84 (4.21)	64.93 (6.28)	−2.57*
IntQ1	47.20 (5.68)	51.54 (6.34)	−2.68*
IntQ5	71.06 (5.06)	75.71 (4.69)	−3.67**
IntSD	8.97 (1.94)	9.69 (1.92)	−2.28*
IntFracRise	0.206 (0.130)	0.225 (0.105)	−0.57 ns
IntFracFall	0.105 (0.077)	0.144 (0.102)	−1.28 ns
Shimmer	0.148 (0.122)	0.100 (0.041)	1.79 ns
F1M	564.18 (90.79)	571.46 (60.39)	−0.35 ns
F2M	1520.0 (163.0)	1454.4 (140.9)	1.52 ns
F3M	2543.9 (130.0)	2516.9 (120.4)	0.73 ns
F1B	151.60 (70.64)	143.08 (56.67)	0.56 ns
F2B	388.50 (225.65)	342.60 (256.37)	1.22 ns
F3B	409.5 (138.1)	361.3 (143.2)	1.82 ns
H1MH2	−8.96 (9.54)	−5.95 (11.13)	−1.54 ns
H1MA3	14.32 (12.07)	15.85 (12.85)	−0.82 ns
SyllableDurM	0.462 (0.165)	0.343 (0.113)	2.51*

Note: $N = 15–17$.
* $p < .05$.
** $p < .01$.

intensity) and lower IntFracRise (percentage of frames with intensity rise) than neutral speech. Resigned utterances, in contrast, had lower IntM, IntQ1 and IntQ5 than neutral speech, and also had smaller IntSD. For the formant cues, the only significant effect was found for F2B (median bandwidth of F2) where irritated speech had higher values than neutral speech. No significant differences were found for the voice source cues. Finally, both irritated and resigned speech had higher values of SyllableDurM (i.e., slower speech rate) than neutral speech. The largest effects were found for F0 and intensity cues.

Generally, the results were similar to those previously obtained using (mainly) posed expressions (see Table 1), though in contrast to most previous studies irritated speech had slower speech rate than neutral speech in our study.[5] It should also be noted that for many acoustic measures no statistically significant differences could be discerned. Unlike many previous studies of authentic expressions, we utilized a within-persons design, where each speaker provided a neutral baseline with which to compare the affective expressions. This design likely increased the power of finding differences. However, though we did reduce the number of acoustic measures to only include those that explained most variance, the set of acoustic measures is still sizable and hence many significance tests need to be performed which may cause alpha-inflation. We followed the recommendations for exploratory research to present uncorrected p-values and to let future research corroborate the findings. However, it should be noted that if we were to use the Bonferroni correction to guard against alpha-inflation, only differences with $p < .002$ would be considered statistically significant, and this way no statistically significant differences would remain for resigned speech (see Table 5).

Initially we had also intended to compare affective speech with high and low emotion intensity within persons. However, expressions of irritation and resignation from the same speaker with different levels of emotion intensity were very scarce in our corpus, so these analyses had to be canceled.

[5] In the present study, irritation may have been associated with hyper-articulation, because many of the utterances originally labeled as *emphatic* were classified as irritated by the listeners. This may have contributed to the finding that irritated utterances were slower than neutral utterances.

6.2.2. Associations between acoustic measures and listeners' perception of affect

Another main research question was to investigate the associations between listeners' perception of affect and acoustic measures. Therefore we calculated the correlations (Pearson *r*) between the acoustic measures and the listeners' mean ratings across all listeners for each utterance, see Table 6. The many significant correlations suggest that listeners utilized many different acoustic cues in judging the emotional content of the utterances. Ratings of irritation, resignation and neutral were correlated with many acoustic measures related to F0 and intensity, whereas fewer correlations were found between the rating scales and the formant and voice source cues. The largest effects were found for the associations between listener ratings and intensity cues. Again, the acoustical correlates of listeners' ratings of irritation and resignation were generally in the same direction as found in previous research on posed expressions (e.g., as shown in Table 1).

Listeners' ratings of *emotion intensity* were also correlated with many acoustic measures (see Table 6). The results suggest that utterances which were perceived as expressing affect with high emotion intensity were associated with high pitch and high intensity. When comparing these results with previous studies on the acoustical correlates of emotion intensity (e.g., Laukka et al., 2005), all statistically significant correlations were in the same direction though the magnitude of the correlations was smaller in the present study. The profiles of acoustic correlates for irritation and emotion intensity turned out to be very similar, which is not surprising considering the high correlation between listeners' ratings on these two scales in the present study.

We also wanted to investigate if the listeners' ratings could be predicted from the acoustical measurements. To this end, we conducted multiple regression analyses with the listeners' mean ratings of irritation, resignation, neutral, and emotion intensity (across all listeners) as dependent variables, and the values of the acoustic measurements for each utterance as independent variables. To avoid including independent variables with high multicollinearity, we excluded the following cues from the analyses: F0M and F0Q5 (highly correlated with F0Q1), IntM (highly correlated with IntQ5), H1MH2 (highly correlated with H1MA3), and Shimmer (highly correlated with Jitter). The remaining 18 acoustic measures were included in the model in a step-wise

Table 6
Correlations (Pearson *r*) between acoustic measures and listeners' mean ratings of irritation, resignation, neutral, and emotion intensity.

Acoustic measure	Irritation	Resignation	Neutral	Emotion intensity
F0M	.19**	−.12 ns	−.15*	.21**
F0Q1	.17*	.11 ns	−.24***	.17*
F0Q5	.14 ns	−.25***	−.03 ns	.16*
F0SD	−.03 ns	−.40***	.22***	.00 ns
F0FracRise	.02 ns	−.29***	.10 ns	.02 ns
F0FracFall	−.15*	−.06 ns	.17*	−.04 ns
Jitter	−.01 ns	.05 ns	−.03 ns	.05 ns
IntM	.44***	−.46***	−.12 ns	.47***
IntQ1	.27***	−.22**	−.08 ns	.30***
IntQ5	.45***	−.58***	−.06 ns	.46***
IntSD	.05 ns	−.23***	.05 ns	.02 ns
IntFracRise	−.22**	−.09 ns	.23***	−.15*
IntFracFall	.04 ns	−.16*	.07 ns	.10 ns
Shimmer	−.02 ns	.04 ns	.01 ns	.05 ns
F1M	.10 ns	.04 ns	−.17*	.18*
F2M	.04 ns	.02 ns	−.10 ns	.09 ns
F3M	−.02	.07 ns	−.07 ns	.00 ns
F1B	.02 ns	.03 ns	−.06 ns	.10 ns
F2B	.18*	.10 ns	−.22**	.23***
F3B	.06 ns	.08 ns	−.11 ns	.08 ns
H1MH2	.01 ns	−.11 ns	.00 ns	.05 ns
H1MA3	−.19**	.01 ns	.13 ns	−.14 ns
SyllableDurM	.20**	.08 ns	−.22**	.15*

Note: $N = 200$.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

fashion, and separate analyses were performed for ratings of irritation, resignation, neutral, and emotion intensity. The included acoustic measures were not highly intercorrelated (maximum correlation $r = -.56$ for F0FracFall/F0FracRise).

The results from the multiple regression analyses are shown in Table 7, and reveal that the linear models could explain around 40–50% of the variance in the listeners ratings of irritation ($R^2 = .39$) and resignation ($R^2 = .54$), but only a third of the variance in ratings of neutral ($R^2 = .32$). The amount of explained variance was similar for emotion intensity as for irritation ($R^2 = .38$). The two most important predictors in terms of variance accounted for were IntQ5 and F0FracRise (percentage of frames with F0 rise) for all scales (except for emotion intensity, for which F2B received a slightly higher beta weight than F0FracRise). However, the exact configurations of these two acoustic measures differed between the scales. While *high* IntQ5 was predictive of high irritation and emotion intensity, *low* IntQ5 was predictive of high resignation and neutral. Also, *low* F0FracRise was predictive of high irritation, resignation and emotion intensity, but *high* F0FracRise was predictive of high ratings of neutral. Thus, F0FracRise seemed to differentiate between neutral and affective utterances in our sample, whereas IntQ5 differentiated between irritated and resigned utterances. The results for the other acoustic cues were less clear, but cues such as Jitter, IntFracRise, and F2B received significant beta weights for several scales, and several other acoustic measures were important predictors for at least one scale (e.g., low F0SD predicted high resignation, and low H1MA3 predicted high irritation).

6.3. Automatic detection of irritation and resignation

The last main research question addressed in the present study was to investigate the automatic detection of irritation and resignation and to compare the performance of the automatic classifiers with human performance. First, we used the results from the previous listening test with 20 judges as criteria for determining which utterances were to be categorized as irritated, resigned or neutral. All utterances which were not clearly categorized as irritated, resigned or neutral were removed ($N = 8$). Also, for the purpose of performing speaker adapted automatic detection, all speakers for whom neutral utterances were not available were removed from the corpus. This procedure left 59 speakers with a total of 139 neutral, 33 irritated and 14 resigned utterances, where 59 of the neutral utterances were reserved for speaker adaptation.[6]

Because the data used for automatic detection were based on human judgments, we wanted to compare the results of automatic detection with those of human perception. We used the leave-one-out method to calculate an estimate of human performance with which to compare the automatic classifiers. By leaving listeners out one at a time, and do majority voting for all others, the detection counts for all left out listeners were accumulated in a confusion matrix. Thus, the reported human performance is an average for all the listeners.

The results from the LDA classification (both with and without speaker adaptation), compared with human performance, are shown in Table 8. We follow the suggestion of Scherer (2003), and report the results in the form of a confusion matrix. Though the analysis of misclassifications is often as informative as the analysis of correct classifications, few previous studies on automatic detection have reported confusion matrices. As shown in Table 8, both classifiers performed close to human performance and with better than chance performance. The average recall (defined as the mean of the diagonal of the confusion matrix), across all three categories, was as follows: 62.3% (LDA, no adaptation), 54.3% (LDA, speaker adaptation), and 57.7% for human performance estimated with the leave-one-out method (the chance level is 33% in a three alternative classification task). Few previous studies on authentic expressions have performed automatic classification of the same categories as the present study (i.e., irritation vs. resignation vs. neutral). A couple of studies have classified speech into angry/frustrated or neutral and report classification performance of around 60–70% recall, which is comparable to the performance in the present study, though we classified the utterances into three categories (e.g., Ang et al., 2002; Burkhardt et al., 2006, 2008). Also, Batliner et al. (2003b) reported around 50% average recall for the categories “no problem” (neutral and joy), helplessness and anger, and

[6] By subtracting the vector mean of one independent neutral utterance per speaker, we hoped to achieve an approximation to using all the neutral data per speaker, which was done in the first part of our study. We also wanted to pursue a quite realistic application-like approach, assuming that the affective detector may be used in a voice based service. In this case, one independent neutral utterance of each user may, for example, be collected for adaptation during enrollment for the service.

Table 7
Summary of results from multiple regression analyses of relationships between acoustic measures and listeners' judgments in terms of multiple correlations (R) and standardized regression coefficients (β). A missing value means that the acoustic feature was not included in the model.

	Irritation	Resignation	Neutral	Emotion intensity
R	$.63^{***}$	$.73^{***}$	$.57^{***}$	$.61^{***}$
Acoustic measure	β			
F0Q1	.13 ns	–	−.14 ns	.10 ns
F0SD	–	$-.30^{***}$	.12 ns	−.01 ns
F0FracRise	$-.31^{***}$	$-.34^{***}$	$.41^{***}$	$-.21^{*}$
F0FracFall	$-.27^{**}$	−.13 ns	$.32^{***}$	−.11 ns
Jitter	$.16^{*}$	.08 ns	$-.19^{**}$	$.17^{**}$
IntQ5	$.54^{***}$	$-.47^{***}$	$-.15^{*}$	$.51^{***}$
IntSD	–	–	−.10 ns	–
IntFracRise	$-.17^{**}$	–	$.16^{*}$	$-.13^{*}$
F1M	–	–	−.10 ns	.08 ns
F3M	$.14^{*}$	–	−.14 ns	.09 ns
F1B	−.06 ns	$.12^{*}$	–	–
F2B	$.13^{*}$	$.12^{*}$	$-.18^{**}$	$.23^{***}$
F3B	.12 ns	.09 ns	−.11 ns	.11 ns
H1MA3	$-.17^{*}$	.09 ns	.10 ns	−.15 ns
SyllableDurM	–	$.13^{*}$	−.11 ns	–

Note: $N = 195–200$.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

in another study, Batliner et al. (2006) reported around 60% average recall for the classification of four classes, including neutral and anger. Regarding resignation, Vidrascu and Devillers (2007) reported that their classifier performed at around 40% recall for sadness using blind features, which is again comparable to the performance in the present study. However, they classified among five different categories and also included lexical cues, so it is difficult to make a direct comparison. It is in general difficult to compare classification results across different studies, because the corpora, classifiers, acoustic measures, and affective classes differ widely between studies. Therefore the above comparisons are best seen as very coarse approximations.

The performance of the automatic classifiers was very similar to that of the listeners. By using one-tailed two-proportion z-tests, we concluded that humans were better than both classifiers at detecting neutral utterances ($p < .0001$ without adaptation, $p < .05$ with adaptation), but not better than either classifier for irritation or resignation. While the LDA classifiers assume uniform prior distribution of expressive categories, the human expectations for this experiment are largely unknown. It is possible that humans assume neutral expressions to be more common than affective expressions, based on their prior experience of everyday speech. If so, this could partly explain the better detection of neutral utterances for humans than for the automatic classifiers. Further, the classifier without speaker adaptation performed slightly better than the one with adaptation – although the difference was not statistically significant. This contrasts with earlier studies reporting better performance for classifiers with speaker adaptation (e.g., Austermann et al., 2005; Grimm et al., 2007). This finding could be due to artifacts of the present corpus, like the fact that the numbers of utterances falling in each expressive category were quite small, or that within-speaker variability made the adaptation (which was based on only one neutral utterance per speaker) unreliable. However, it may also be the case that the speaker dependent component of the acoustic measurements was not as prominent in the present corpus.

As shown in Table 8, the confusion matrices for the classifiers and human listeners were very similar. Both machines and humans seldom confused irritation and resignation, but neutral utterances were quite frequently misclassified as resigned. This suggests that the human judges may have used inference mechanisms that are comparable to optimized statistical classification procedures, as suggested by Banse and Scherer (1996). However, when we calculated the average inter-rater reliability (in terms of Cohen's kappa statistic) between each

Table 8
Confusion matrices for the detection of irritation, resignation, and neutral for (a) automatic detection using LDA (both with and without speaker adaptation) and (b) human perception.

		Recall (%)		
		Irritation	Resignation	Neutral
No adaptation	Irritation	**69.7**	6.1	24.2
	Resignation	0.0	**64.3**	35.7
	Neutral	28.7	21.2	**50.0**
Adaptation	Irritation	**57.6**	9.1	33.3
	Resignation	14.3	**42.9**	42.9
	Neutral	16.3	21.2	**62.5**
Human performance	Irritation	**56.7**	7.4	35.9
	Resignation	18.6	**45.3**	36.0
	Neutral	15.5	13.3	**71.2**

listener and each machine, the mean kappa values suggested only slight agreement: listeners/LDA (no adaptation), $\kappa = .19$; listeners/LDA (with speaker adaptation), $\kappa = .16$. Thus, the human listeners and the machines did not necessarily classify individual utterances similarly, even though the overall performances are comparable.

Previous studies have reported that vocal expressions with high emotion intensity are easier to recognize than expressions with low emotion intensity (e.g., Juslin and Laukka, 2001). Therefore we compared the emotional intensity, as judged by the listeners, of correctly and wrongly classified utterances for both classifiers (using *t*-tests). The results showed that wrongly classified utterances had significantly lower emotion intensity than correctly classified utterances for irritated speech, but this difference was only significant for the non-adaptive system ($p < .0001$). No significant differences in emotion intensity were found for resigned or neutral utterances. Thus the finding that vocal expressions with high emotion intensity should be easier to recognize could only be partially replicated in the present study using authentic speech and automatic classification of affect.

Finally, the rankings of each acoustic measure are shown in Table 9. A comparison of the two lists provide information about which measures are speaker dependent, and which cues can be used for classification with-

Table 9
The ranking of each acoustic measure used in the LDA classifications (both with and without speaker adaptation).

No adaptation		Adaptation	
Rank	Acoustic measure	Rank	Acoustic measure
1.00	IntQ5	0.63	F0Q5
0.56	IntFracRise	0.56	SyllableDurM
0.53	F1M	0.44	IntQ5
0.46	F3M	0.40	IntSD
0.41	F0FracRise	0.35	F0FracFall
0.36	H1MH2	0.34	F3M
0.29	F3B	0.33	IntM
0.27	F0Q5	0.28	F1B
0.20	IntM	0.27	IntQ1
0.19	IntFracFall	0.26	F0Q1
0.19	F2M	0.25	FracRise
0.19	F0M	0.22	Shimmer
0.15	F0Q1	0.22	IntFracRise
0.13	H1MA3	0.20	F2B
0.11	Jitter	0.15	H1MA3
0.07	F1B	0.11	Jitter
0.04	IntQ1	0.07	H1MH2
0.03	IntSD	0.04	F0SD
0.03	F2M		

out speaker adaptation. As expected, different acoustic measures featured on the lists for the two classifiers. For the classifier with no speaker adaptation, IntQ5 received the highest ranking followed by IntFracRise, F1M, and F3M. For the classifier with speaker adaptation, however, F0Q5 was the highest ranking acoustic measure, followed by SyllableDurM, IntQ5 and IntSD. Pitch cues are highly speaker dependent, so it was not surprising to find that F0Q5 was highly ranked by the speaker dependent classifier but was not as important for the classifier without speaker adaptation. The results also suggest that SyllableDurM is an important cue for classification of affect when individual differences are controlled for, whereas IntQ5 worked well both with and without speaker adaptation. It should be noted that lower ranked features are less reliable than the higher ranked ones.[7] When inspecting the five top measures in the ranked lists presented in Table 9, all of these measures (except for F3M) also received significant effects in the within-speaker analyses between affective and neutral speech (see Table 4), and/or in the correlation analyses between listener ratings and acoustic measures (see Table 6). Thus, these acoustic measures can be considered important predictors of affect in authentic speech.

7. Discussion

The main findings of the present study may be summarized as follows. Utilizing a speech corpus consisting of authentic affective speech from human–computer interactions recorded from call-centers, we first found several significant acoustic differences between speech perceived as neutral and speech perceived as irritated or resigned, using a within-persons design. Second, listeners' ratings of irritation, resignation, neutral and emotion intensity were associated with acoustic features related to pitch, intensity, formants, voice source, and temporal characteristics of speech. Third, automatic detection using LDA classifiers (both with and without speaker adaptation) performed at a level comparable to human performance in classifying the utterances into irritated, resigned and neutral. Fourth, clearly perceived exemplars of irritation and resignation were rare in our corpus. The implications of these findings are discussed in the following sections.

For the "classic" acoustic measures related to pitch, intensity and durations, the results corroborate earlier findings obtained with mainly posed expressions (see Juslin and Laukka, 2003). For example, irritation was associated with higher pitch (F0M, F0Q1, and F0Q5) and intensity (IntM and IntQ5) than neutral speech, whereas resignation was associated with lowered pitch (F0M and F0Q5) and intensity (IntM, IntQ1, and IntQ5). Also, the variability of pitch and intensity (F0SD and IntSD) was smaller for resigned speech than for neutral speech, and resigned utterances also had lower speech rate (i.e., higher SyllableDurM) than neutral utterances. Our battery of acoustic cues, however, also included cues that have seldom been included in studies of vocal expressions, like formants and voice source cues. The results regarding formant cues were not very clear, but suggest that F2B (median bandwidth of F2) could potentially co-vary with affect (e.g., irritated utterances showed higher F2B than neutral speech). However, because the semantic content of each utterance was different, and because formants largely determine vowel quality, it is possible that this finding is an artifact of the semantic content of affective versus neutral utterances. Similarly, the lack of control over the semantic content may have obscured possible effects of irritation and resignation on the voice source cues. As the results stand, the only significant effect was a negative correlation between H1MA3 (a measure of spectral tilt at the higher formant frequencies) and listeners' ratings of irritation. Because H1MA3 is expected to be large for breathy voices and small for creaky voices (e.g., Hanson, 1997), this correlation suggests an association between perceived irritation and creaky voice quality. This finding is in accord with previous studies conducted on posed expressions (e.g., Gendrot, 2003; Laukkanen et al., 1996; Yanushevskaya et al., 2007; Yuan et al., 2002), but more studies need to be conducted on voice source measures before stronger conclusions are drawn. Also, it should be noted that most previous studies have looked at slightly different operationalizations of breathiness, or utilized inverse filtering or EGG methods whereas we automatically measured spectral correlates of voice source features. Additionally, concerning other more rarely investigated acoustic measures our

[7] The sequential forward search wrapper algorithm used in the present study has been shown to be susceptible to nesting effects (for a recent discussion, see Ververidis and Kotropoulos, 2008). However, more important (higher ranked) features are less prone to nesting effects than the less important ones. Also, nesting effects in terms of the rank of each feature are canceled out to some extent by the averaging of rank throughout the cross-validation sets.

results suggested that jitter (small scale perturbations in F0) was associated with perceived irritation in accordance with some previous studies conducted on posed expressions (see Juslin and Laukka, 2003). Finally, our study also suggested that the proportions of frames with F0 or intensity rise or fall (F0FracRise/Fall and IntFracRise/Fall) were associated with perceived irritation and resignation, but more studies are needed to confirm these observations.

A comparison of the acoustical correlates of authentic irritation and resignation, and the acoustical correlates of (mainly) posed anger and sadness (as summarized in Table 1), reveals that all significant effects go in the same direction (except for irritated speech which had slower speech rate than neutral speech in our study, contrary to most previous studies). Thus, the present results mostly agree with both: (a) previous results from studies of posed expressions and (b) Scherer's (1986) predictions based on somatic alterations associated with emotional responses. Therefore the present results are consistent with the idea that posed expressions are at least partly based on spontaneous expressions, which in turn are at least partly based on physiological emotional responses (see also Laukka et al., 2008). However, the effect sizes in the present study, though statistically significant, were much smaller than in comparable studies that have used posed expressions (see also Williams and Stevens, 1972). This supports the view that spontaneous expressions are less prototypical and more subtle than those often used in studies of portrayed expressions, as suggested by Cowie and Cornelius (2003) (see also Cowie, 2009). However, the small effect sizes could also partly result from the lack of control over the semantic content of the utterances in the present study, which likely did introduce extra variability in the acoustic data.

In many studies of spontaneous expressions, it is difficult to determine the affective states of the speakers with any certainty, but the context of the recordings allowed us to infer the speakers' probable affective states with reasonable certainty in the present study. For example, for both irritation and resignation the speakers' goals were obstructed (i.e., he or she could not get their message through). The irritated utterances often appeared in the beginning of the speech episodes, where the obstruction was novel and the speakers were still certain that they would be understood in the end if they just kept on trying. Resignation instead often occurred at the end of the speech episodes, where the speakers had realized that they would not be understood this time, and that there was nothing they could do about it. The labeling of affective states is very tricky and there is no general agreement on what different labels stand for. We chose to call the affective states under study irritation and resignation, because the context clearly placed them in the anger and sadness families (though the affective states in our corpus were much less intense than the full-blown emotions anger and sadness). As it turned out, the acoustic correlates for authentic irritation and resignation were very similar to those of anger and sadness, which is what one would expect if affect in speech was coded in terms of broad emotion categories/families (Laukka, 2005). However, it should be noted that the view that affect in speech is primarily coded in terms of broad emotion categories/families does not exclude the possibility that listeners can also recognize differences within these categories. Indeed, previous research has shown that listeners can recognize between emotion portrayals with weak and strong emotion intensity (e.g., Juslin and Laukka, 2001) and between more and less prototypical expressions of the same emotion categories (e.g., Laukka et al., 2009).

The comparison of human and machine classification performance gave at hand that overall the automatic classifiers performed at a level comparable with humans, and the resulting confusion matrices were also similar for human listeners and machines. However, the human listeners and the machines did not necessarily classify individual utterances similarly. This pattern of results suggests that, while the listeners' classification judgments may bear some resemblance with optimized statistical classification methods, statistical methods (or at least the statistical method used in this study, LDA) do not provide an exact model of human inference processes (see also Banse and Scherer, 1996). It could for example be the case that human listeners base their inferences on an ideal, or prototypical representation, of each expression (Laukka et al., 2009). It is also possible that humans have several different prototypes per expression that they have acquired through their experience. In any case it can be argued that humans have a more representative sampling of instances of vocal expressions occurring in everyday life than the machines, which have to base their inference procedures on only those instances that were encountered in the training phase. It could also be hypothesized that human listeners may utilize speaker dependent acoustic information, because they may have internal representations of how different types of voices usually sound in various situations, whereas this would be difficult for many automatic detection applications. The fact that automatic classifiers are highly specialized at classifying among a very specific set of utterances could account for the fact that it is difficult to achieve high recall when using an automatic classifier on another speech corpus than what it was originally trained on. To date, most studies of automatic detection have been content with acquiring acceptably high recognition accuracy, and have not tried to more closely model human performance. It has been shown that classifiers which have been trained on many corpora of different kinds of speech (including both posed and authentic expressions) become more robust and can handle more varied expressions of each emotion class (Shami and Verhelst, 2007). It remains an intriguing question for future research to investigate whether classifiers that are trained on many different corpora also more closely model human classification performance. Further, it may also be the case that listeners and machines utilized somewhat different acoustic cues in order to reach their decisions. On the one hand human listeners were able to utilize features such as intonation patterns, sentence stress and locations of pauses, which were inaccessible to the automatic classifiers, and on the other hand the perceptual correlates of many of the acoustic features available to the classifiers are unknown.

The present study also has several limitations, which may limit the generalizability of the findings. First and most importantly our corpus included rather small numbers of stimuli belonging to each expressive category which may have interfered with the robustness of the LDA analyses, especially concerning the automatic detection of resignation. The low percentage of affective material is a characteristic that our database shares with many other databases of spontaneous behavior (e.g., Ang et al., 2002; Grimm et al., 2008; Vogt and André, 2005). Affect in speech may represent various classes of affective phenomena ranging from intense full-blown emotions to milder affective states like moods, stances, and attitudes (Scherer, 2003). Because intense emotions occur less frequently than mild and moderately intense affective states in our daily lives, expressions of intense states should consequently also occur less frequently than expressions of milder states (see Scherer et al. (2004), for data on the frequencies with which various emotions may occur in everyday life). Therefore expressions of full-blown emotions may represent only the tip of the iceberg, whereas the iceberg itself consists of milder and more subtle affective states.[8] The results from the present study add nuance to this picture by suggesting that *clearly perceived* exemplars of also milder affective states may be rare in spontaneous behavior. A tentative interpretation of this finding is that clearly perceived expressions of milder affective states are rare because they need to bear some resemblance to prototypical expressions in order to be clearly perceived (as suggested by the present results), and that the combination of push and pull effects (i.e., people often strategically pose or try to mask their expressions) render prototypical expressions rare in spontaneous interactions. However, it could also be the case that most corpora of naturalistic affective speech (including ours) do not contain enough clearly perceived exemplars simply because the interactions that the speech has been collected from have not been emotional enough.

Our study also had other limitations. For example, the lexical content of the speech was not controlled which may have influenced both the listeners' recognition of affect and the acoustic measurements (e.g., formant and voice source cues), and probably added extra variability (i.e., variability not due to the vocal expression) in the acoustic measurements. Also, the utterances in our corpus consisted of telephone speech and therefore included a more limited vocabulary, consisting of mainly single words and short phrases, than ordinary spontaneous speech. When talking with voice controlled telephone services people often do not feel the pressure to follow conventional politeness rules. The lack of a human counterpart in the dialogues may also have biased our vocal expressions toward resulting from push effects, rather than pull effects. Further, the speech signal was of low bandwidth and disturbed by a noisy environment which may have interfered with the acoustic feature extraction. However, there are also many applications for telephone speech, for example in developing call center applications that can recognize and respond to the user's vocal expressions. Moreover, our corpus only included two different expressive categories (i.e., irritation and resignation) whereas everyday expressions may be much more varied. Finally, we would like to mention the limitation that this study only concerned the speech modality, whereas in spontaneous conversations affective information is communicated in a multimodal fashion including the voice (both what is said and how it is said) in combination

[8] This is not to downgrade the all important tip of the iceberg. When prototypical expressions *do* occur in spontaneous conversations they are likely to efficiently capture the attention of the perceiver and are thus important for the regulation of social interactions. Also, the study of prototypical expressions warrants interest for theoretical reasons (e.g., for answering questions like "What expressions can be reliably recognized from vocal expressions?").

with facial expressions and bodily gestures (e.g., Bänziger et al., 2006; Barkhuysen et al., 2005; Martin et al., 2006; Zeng et al., 2009). The study of how expression is coordinated in several modalities adds to the complexity of the studies, but is in the end necessary for a more complete understanding of human affect expression.

To conclude, designing and conducting a study of authentic vocal expressions is an undertaking that is fraught with difficulties, especially concerning the collection of emotional corpora. However, such undertakings are necessary to increase the ecological validity of vocal expression studies. To achieve a more complete understanding of affective speech in real-life settings, future research could try out the following avenues of inquiry. First, researchers who wish to study the tip of the iceberg (i.e., clearly perceived expressions) should study interactions that have been carefully designed to ensure that the participants really do engage in interaction with the intention to communicate non-verbally (and that they do not try to suppress their expressions). The design of such studies will require much planning and effort, but will probably increase the ratio of affective versus non-affective speech in the resulting corpora. Second, if the aim instead is to capture the essence of expression in spontaneous conversational speech, we need an improved way of describing the milder and more subtle expressions that make up the bulk of the iceberg. Affective states that could be included in the annotation of naturalistic data include stress, attitudes (e.g., friendly, disliking), states that are partly cognitive and partly emotional (e.g., thoughtful, sure), and mixed affective states (e.g., both angry and sad at the same time; see Douglas-Cowie et al., 2007). Efforts also need to be directed towards investigating the frequencies with which expressions of various types of affective phenomena appear in everyday spontaneous interactions. It remains a challenge for future research to map out the kinds of affects that are expressed in different contexts and situations. Third, because clearly perceived expressions are scarce, researchers should also try to use more efficient ways of annotating their spontaneous data (rather than dividing the expressions into mutually exclusive categories). For example, the use of rating scales, either in the form of multiple category labels or dimensional representations (e.g., activation, valence, potency/dominance/control, and emotion intensity), allows a more fine-grained description of the perceived affect that can account for expressions that are not readily placed in any one category (e.g., Laukka et al., 2005). The use of free responses would also give much needed information about what labels people spontaneously use to describe the range of percepts that they form when decoding spontaneous expressions (see Cowie, 2009). Finally, future studies on the acoustic correlates of affective states could also benefit from the machine learning based methods of handling large amounts of acoustical data used in studies of automatic detection. Likewise, automatic detection methods could profitably be used for modeling human classification performance, and could in that way be used to study the inference processes involved in human emotion recognition.

Acknowledgements

The research was supported by the Swedish Research Council and SYLFF through grants to the first author. Parts of the work were performed within CHIL (Computers in the Human Interaction Loop), an EU 6th Framework IP (506909). We would like to thank Voice Provider for providing the speech material. Petri Laukka is currently at the Department of Psychology, Stockholm University, Stockholm, Sweden.

References

Ang, J., Dhillon, R., Krupski, A., Shriberg, E., Stolcke, A., 2002. Prosody-based automatic detection of annoyance and frustration in human–computer dialog. In: Proceedings of the 7th International Conference on Spoken Language Processing, Denver, CO, pp. 2037–2040.

Athanaselis, T., Bakamidis, S., Dologlou, I., Cowie, R., Douglas-Cowie, R., Cox, C., 2005. ASR for emotional speech: clarifying the issues and enhancing performance. Neural Netw. 18, 437–444.

Aubergé, V., Audibert, N., Rilliard, A., 2006. De E-Wiz à C-Clone: Recueil, modélisation et synthèse d'expressions authentiques. Revue d'Intelligence Artificielle 20, 499–527.

Austermann, A., Esau, N., Kleinjohann, L., Kleinjohann, B., 2005. Fuzzy emotion recognition in natural speech dialogue. In: Proceedings of the 2005 IEEE International Workshop on Robots and Human Interactive Communication, pp. 317–322.

Bachorowski, J.-A., Owren, M.J., 1995. Vocal expression of emotion: acoustical properties of speech are associated with emotional intensity and context. Psychol. Sci. 6, 219–224.

Banse, R., Scherer, K.R., 1996. Acoustic profiles in vocal emotion expression. J. Pers. Soc. Psychol. 70, 614–636.

Bänziger, T., Pirker, H., Scherer, K.R., 2006. Gemep – Geneva multimodal emotion portrayals: a corpus for the study of multimodal emotional expressions. In: Proceedings of the LREC 2006 Workshop on Corpora for Research on Emotion and Affect, Genoa, Italy, pp. 15–19.

Barkhuysen, P., Krahmer, E., Swerts, M., 2005. Problem detection in human–machine interactions based on facial expressions of users. Speech Commun. 45, 343–359.

Barrett, J., Paus, T., 2002. Affect-induced changes in speech production. Exp. Brain Res. 146, 531–537.

Batliner, A., Fischer, K., Huber, R., Spilker, J., Nöth, E., 2003a. How to find trouble in communication. Speech Commun. 40, 117–143.

Batliner, A., Zeissler, V., Frank, C., Adelhardt, J., Shi, R.P., Nöth, E., 2003b. We are not amused – but how do you know? User states in a multi-modal dialogue system. In: Proceedings of the 8th European Conference on Speech Communication and Technology, Geneva, Switzerland, pp. 733–736.

Batliner, A., Steidl, S., Schuller, B., Seppi, D., Laskowski, K., Vogt, T., Devillers, L., Vidrascu, L., Amir, N., Kessous, L., Aharonson, V., 2006. Combining efforts for improving automatic classification of emotional user states. In: Proceedings of the 5th Slovenian and 1st International Language Technologies Conference, Ljubljana, Slovenia, pp. 240–245.

Batliner, A., Steidl, S., Hacker, C., Nöth, E., 2008. Private emotions versus social interaction: a data-driven approach towards analysing emotion in speech. User Model. User-Adap. Inter. 18, 175–206.

Boersma, P., Weenink, D., 2007. Praat: doing phonetics by computer. Institute of Phonetic Sciences. University of Amsterdam, Amsterdam, The Netherlands (computer program).

Bonner, M.R., 1943. Changes in the speech pattern under emotional tension. Am. J. Psychol. 56, 262–273.

Breazeal, C., Aryananda, L., 2002. Recognition of affective communicative intent in robot-directed speech. Auton. Robot. 12, 83–104.

Brehm, J., 1999. The intensity of emotion. Pers. Soc. Psychol. Rev. 3, 2–22.

Buck, R., 1984. The Communication of Emotion. Guilford Press, New York.

Burkhardt, F., Ajmera, J., Englert, R., Stegmann, J., Burleson, W., 2006. Detecting anger in automated voice portal dialogs. In: Proceedings of the 9th International Conference on Spoken Language Processing, Pittsburgh, PA (paper 1977-Tue2A3O.3).

Burkhardt, F., Huber, R., Stegmann, J., 2008. Advances in anger detection with real life data. In: Tagungsband 19. Konferenz Elektronische Sprachsignalverarbeitung (ESSV 2008), Frankfurt, Germany.

Campbell, N., 2005. Getting to the heart of the matter: speech as the expression of affect; rather than just text or language. Lang. Resour. Eval. 39, 109–118.

Chateau, N., Maffiolo, V., Blouin, C., 2004. Analysis of emotional speech in voice mail messages: the influence of speakers' gender. In: Proceedings of the 8th International Conference on Spoken Language Processing, Jeju Island, Korea, pp. 885–888.

Cowie, R., 2009. Perceiving emotion: towards a realistic understanding of the task. Philos. Trans. Roy. Soc. B 364, 3515–3525.

Cowie, R., Cornelius, R.R., 2003. Describing the emotional states that are expressed in speech. Speech Commun. 40, 5–32.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.G., 2001. Emotion recognition in human–computer interaction. IEEE Signal Process. Mag. 18 (1), 32–80.

Darwin, C., 1872/1998. The Expression of the Emotions in Man and Animals. Oxford University Press, New York.

Davitz, J.R., 1964. A review of research concerned with facial and vocal expressions of emotion. In: Davitz, J.R. (Ed.), The Communication of Emotional Meaning. McGraw-Hill, New York, pp. 13–29.

Dellaert, F., Polzin, T., Waibel, A., 1996. Recognizing emotion in speech. In: Proceedings of the 4th International Conference on Spoken Language Processing, Philadelphia, PA, pp. 1970–1973.

Devillers, L., Vidrascu, L., Lamel, L., 2005. Challenges in real-life emotion annotation and machine learning based detection. Neural Netw. 18, 407–422.

Douglas-Cowie, E., Campbell, N., Cowie, R., Roach, P., 2003. Emotional speech: towards a new generation of databases. Speech Commun. 40, 33–60.

Douglas-Cowie, E., Cowie, R., Sneddon, I., Cox, C., Lowry, O., McRorie, M., Martin, J.-C., Devillers, L., Abrilian, S., Batliner, A., Amir, N., Karpouzis, K., 2007. The HUMAINE database: addressing the collection and annotation of naturalistic and induced emotional data. In: Proceedings of the 2nd International Conference on Affective Computing and Intelligent Interaction, Lisbon, Portugal, pp. 488–500.

Ekman, P., 1992. An argument for basic emotions. Cogn. Emotion 6, 169–200.

Ekman, P., 2003. Emotions Revealed. Henry Holt, New York.

Ekman, P., Friesen, W.V., 1969. The repertoire of nonverbal behavior: categories, origins, usage, and coding. Semiotica 1, 49–98.

Eldred, S.H., Price, D.B., 1958. A linguistic evaluation of feeling states in psychotherapy. Psychiatry 21, 115–121.

Forsell, M., 2007. Acoustic Correlates of Perceived Emotions in Speech. Master thesis, Department of Speech, Music and Hearing, KTH, Stockholm, Sweden.

Frick, R.W., 1986. The prosodic expression of anger: differentiating threat and frustration. Aggressive Behav. 12, 121–128.

Fukunaga, K., 1990. Introduction to Statistical Pattern Recognition. Academic Press, San Diego, CA.

Gendrot, C., 2003. Rôle de la qualité de la voix dans la simulation des émotions: Une étude perceptive et physiologique. Revue Parole 27, 137–158.

Greasley, P., Sherrard, C., Waterman, M., 2000. Emotion in language and speech: methodological issues in naturalistic settings. Lang. Speech 43, 355–375.

Grimm, M., Kroschel, K., Mower, E., Narayanan, S., 2007. Primitives-based evaluation and estimation of emotions in speech. Speech Commun. 49, 787–800.

Grimm, M., Kroschel, K., Narayanan, S., 2008. The Vera am Mittag German audio–visual emotional database. In: Proceedings of the 2008 IEEE International Conference on Multimedia and Expo, Hannover, Germany, pp. 865–868.

Hair, J.F., Anderson, R.E., Tatham, R.L., Black, W.C., 1998. Multivariate Data Analysis, fifth ed. Prentice-Hall, Inc., Upper Saddle River, NJ.

Hall, J.A., 2006. Gender differences in nonverbal communication: similarities, differences, stereotypes, and origins. In: Manusov, V.L., Patterson, M.L. (Eds.), The Handbook of Nonverbal Communication. Sage, Thousand Oaks, CA, pp. 201–218.

Hanson, H.M., 1997. Glottal characteristics of female speakers: acoustic correlates. J. Acoust. Soc. Am. 101, 466–481.

Johnstone, T., van Reekum, C.M., Hird, K., Kirsner, K., Scherer, K.R., 2005. Affective speech elicited with a computer game. Emotion 5, 513–518.

Johnstone, T., van Reekum, C.M., Bänziger, T., Hird, K., Kirsner, K., Scherer, K.R., 2007. The effects of difficulty and gain versus loss on vocal physiology and acoustics. Psychophysiology 44, 827–837.

Juslin, P.N., Laukka, P., 2001. Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion. Emotion 1, 381–412.

Juslin, P.N., Laukka, P., 2003. Communication of emotions in vocal expression and music performance: different channels, same code? Psychol. Bull. 129, 770–814.

Kandali, A.B., Routray, A., Basu, T.K., 2009. Vocal emotion recognition in five native languages of Assam using new wavelet features. Int. J. Speech Technol. 12, 1–13.

Laukka, P., 2005. Categorical perception of vocal emotion expressions. Emotion 5, 277–295.

Laukka, P., 2008. Research on vocal expression of emotion: state of the art and future directions. In: Izdebski, K. (Ed.), Emotions in the Human Voice, Foundations, vol. 1. Plural Publishing, San Diego, CA, pp. 153–169.

Laukka, P., Juslin, P.N., 2007. Similar patterns of age-related differences in emotion recognition from speech and music. Motiv. Emotion 31, 182–191.

Laukka, P., Juslin, P.N., Bresin, R., 2005. A dimensional approach to vocal expression of emotion. Cogn. Emotion 19, 633–653.

Laukka, P., Linnman, C., Åhs, F., Pissiota, A., Frans, Ö., Faria, V., Michelgård, Å., Appel, L., Fredrikson, M., Furmark, T., 2008. In a nervous voice: acoustic analysis and perception of anxiety in social phobics' speech. J. Nonverbal Behav. 32, 195–214.

Laukka, P., Audibert, N., Aubergé, V., 2009. Exploring the graded structure of vocal emotion expressions. In: Hancil, S. (Ed.), The Role of Prosody in Affective Speech. Peter Lang, Bern, Switzerland, pp. 241–258.

Laukkanen, A.-M., Vilkman, E., Alku, P., Oksanen, H., 1996. Physical variations related to stress and emotional state: a preliminary study. J. Phonetics 24, 313–335.

Lee, C.M., Narayanan, S.S., 2005. Toward detecting emotions in spoken dialogs. IEEE Trans. Speech Audio Process. 13, 293–303.

Litman, D.J., Forbes-Riley, K., 2006. Recognizing student emotions and attitudes on the basis of utterances in spoken tutoring dialogues with both human and computer tutors. Speech Commun. 48, 559–590.

Martin, J.-C., Niewiadomski, R., Devillers, L., Buisine, S., Pelachaud, C., 2006. Multimodal complex emotions: gesture expressivity and blended facial expressions. Int. J. Humanoid Rob. 3, 269–291.

Morrison, D., Wang, R., De Silva, L.C., 2007. Ensemble methods for spoken emotion recognition in call-centres. Speech Commun. 49, 98–112.

Murray, I.R., Arnott, J.L., 2008. Applying an analysis of acted vocal emotions to improve the simulation of synthetic speech. Comput. Speech Lang. 22, 107–129.

Neiberg, D., Elenius, K., 2008. Automatic recognition of anger in spontaneous speech. In: Proceedings of the Interspeech 2008, Brisbane, Australia, pp. 2755–2758.

Neiberg, D., Elenius, K., Laskowski, K., 2006. Emotion recognition in spontaneous speech using GMMs. In: Proceedings of the 9th International Conference on Spoken Language Processing, Pittsburgh, PA (paper 1581-Tue1A3O.5).

Nicholson, J., Takahashi, K., Nakatsu, R., 2000. Emotion recognition in speech using neural networks. Neural Comput. Appl. 9, 290–296.

Nogueiras, A., Moreno, A., Bonafonte, A., Marino, J.B., 2001. Speech emotion recognition using hidden Markov models. In: Proceedings of the 7th European Conference on Speech Communication and Technology, Aalborg, Denmark, pp. 2679–2682.

Nwe, T.L., Foo, S.W., De Silva, L.C., 2003. Speech emotion recognition using hidden Markov models. Speech Commun. 41, 603–623.

Oudeyer, P.-Y., 2003. The production and recognition of emotions in speech: features and algorithms. Int. J. Hum–Comput. Stud. 59, 157–183.

Petrushin, V.A., 1999. Emotion in speech: recognition and application to call centers. In: Proceedings of the 1999 Conference on Artificial Neural Networks in Engineering (ANNIE'99), St. Louis, MO.

Russell, J.A., Bachorowski, J.-A., Fernandez-Dols, J.-M., 2003. Facial and vocal expressions of emotion. Annu. Rev. Psychol. 54, 329–349.

Scherer, K.R., 1986. Vocal affect expression: a review and a model for future research. Psychol. Bull. 99, 143–165.

Scherer, K.R., 1989. Vocal correlates of emotional arousal and affective disturbance. In: Wagner, H., Manstead, A. (Eds.), Handbook of Social Psychophysiology. Wiley, New York, pp. 165–197.

Scherer, K.R., 2003. Vocal communication of emotion: a review of research paradigms. Speech Commun. 40, 227–256.

Scherer, K.R., Ceschi, G., 2000. Criteria for emotion recognition from verbal and nonverbal expression: studying baggage loss in the airport. Pers. Soc. Psychol. Bull. 26, 327–339.

Scherer, K.R., Tran, V., 2001. Effects of emotion on the process of organisational learning. In: Antal, A.B., Child, J., Dierkes, M., Nonaka, I. (Eds.), Handbook of Organizational Learning and Knowledge. Oxford University Press, New York, pp. 369–392.

Scherer, K.R., Wranik, T., Sangsue, J., Tran, V., Scherer, U., 2004. Emotions in everyday life: probability of occurrence, rick factors, appraisal and reaction patterns. Soc. Sci. Inform. 43, 499–570.

Schröder, M., 2001. Emotional speech synthesis: a review. In: Proceedings of the 7th European Conference on Speech Communication and Technology, Aalborg, Denmark, pp. 561–564.

Schuller, B., Rigoll, G., 2006. Timing levels in segment-based speech emotion recognition. In: Proceedings of the 9th International Conference on Spoken Language Processing, Pittsburgh, PA (paper 1695-Wed2BuP.8).

Schuller, B., Müller, R., Hörnler, B., Höthker, A., Konosu, H., Rigoll, G., 2007. Audiovisual recognition of spontaneous interest within conversations. In: Proceedings of the 9th International Conference on Multimodal Interfaces, Nagoya, Japan, pp. 30–37.

Selting, M., 1994. Emphatic speech style – with special focus on the prosodic signalling of heightened emotive involvement in conversation. J. Pragmat. 22, 375–408.

Shami, M., Verhelst, W., 2007. An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech. Speech Commun. 49, 201–212.

Slaney, M., McRoberts, G., 2003. BabyEars: a recognition system for affective vocalizations. Speech Commun. 39, 367–384.

Sonnemans, J., Frijda, N.H., 1994. The structure of subjective emotional intensity. Cogn. Emotion 8, 329–350.

Tatham, M., Morton, K., 2004. Expression in Speech: Analysis and Synthesis. Oxford University Press, New York.

Toivanen, J., Väyrynen, E., Seppänen, T., 2004. Automatic discrimination of emotion from Finnish. Lang. Speech 47, 383–412.

Touboul, E., 2003. Recognition of Emotions: Automatic Extraction of Spectral Correlates from the Glottal Source. Master thesis, Department of Speech, Music and Hearing, KTH, Stockholm, Sweden.

Ververidis, D., Kotropoulos, C., 2006. Emotional speech recognition: resources, features, and methods. Speech Commun. 48, 1162–1181.

Ververidis, D., Kotropoulos, C., 2008. Fast and accurate sequential floating forward feature selection with the Bayes classifier applied to speech emotion recognition. Signal Process. 88, 2956–2970.

Vidrascu, L., Devillers, L., 2007. Five emotion classes detection in real-world call center data: the use of various types of paralinguistic information. In: Proceedings of the International Workshop on Paralinguistic Speech – between Models and Data (ParaLing'07), Saarbrücken, Germany, pp. 11–16.

Vogt, T., André, E., 2005. Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. In: Proceedings of the 2005 IEEE International Conference on Multimedia and Expo, Amsterdam, The Netherlands, pp. 474–477.

Williams, C.E., Stevens, K.N., 1972. Emotions and speech: some acoustical correlates. J. Acoust. Soc. Am. 52, 1238–1250.

Yanushevskaya, I., Tooher, M., Gobl, C., Ní Chasaide, A., 2007. Time- and amplitude-based voice source correlates of emotional portrayals. In: Proceedings of the 2nd International Conference on Affective Computing and Intelligent Interaction, Lisbon, Portugal, pp. 159–170.

Yuan, J., Shen, L., Chen, F., 2002. The acoustic realization of anger, fear, joy and sadness in Chinese. In: Proceedings of the 7th International Conference on Spoken Language Processing, Denver, CO, pp. 2025–2028.

Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S., 2009. A survey of affect recognition methods: audio, visual, and spontaneous expressions. IEEE Trans. Pattern. Anal. Mach. Intell. 31, 39–58.

Zwick, W.R., Velicer, W.F., 1986. Comparison of five rules for determining the number of components to retain. Psychol. Bull. 99, 432–442.