ARTICLE IN PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT

Speech Communication xxx (2006) xxx–xxx

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Emotional speech recognition: Resources, features, and methods

Dimitrios Ververidis, Constantine Kotropoulos *

Artificial Intelligence and Information Analysis Laboratory, Department of Informatics, Aristotle University of Thessaloniki, University Campus, Box 451, Thessaloniki 541 24, Greece

Received 15 July 2004; received in revised form 19 April 2006; accepted 24 April 2006

Abstract

In this paper we overview emotional speech recognition having in mind three goals. The first goal is to provide an up-to-date record of the available emotional speech data collections. The number of emotional states, the language, the number of speakers, and the kind of speech are briefly addressed. The second goal is to present the most frequent acoustic features used for emotional speech recognition and to assess how the emotion affects them. Typical features are the pitch, the formants, the vocal tract cross-section areas, the mel-frequency cepstral coefficients, the Teager energy operator-based features, the intensity of the speech signal, and the speech rate. The third goal is to review appropriate techniques in order to classify speech into emotional states. We examine separately classification techniques that exploit timing information from which that ignore it. Classification techniques based on hidden Markov models, artificial neural networks, linear discriminant analysis, *k*-nearest neighbors, support vector machines are reviewed.

Keywords: Emotions; Emotional speech data collections; Emotional speech classification; Stress; Interfaces; Acoustic features

1. Introduction

Emotional speech recognition aims at automatically identifying the emotional or physical state of a human being from his or her voice. The emotional and physical states of a speaker are known as emotional aspects of speech and are included in the so-called paralinguistic aspects. Although the emotional state does not alter the linguistic content, it is an important factor in human communication, because it provides feedback information in many applications as it is outlined next.

Making a machine to recognize emotions from speech is not a new idea. The first investigations were conducted around the mid-1980s using statistical properties of certain acoustic features (Van Bezooijen, 1984; Tolkmitt and Scherer, 1986). Ten years later, the evolution of computer architectures made the implementation of more complicated emotion recognition algorithms feasible. Market requirements for automatic services motivate further research. In environments like aircraft cockpits, speech recognition systems were trained by employing stressed speech instead of neutral (Hansen and Cairns, 1995). The acoustic features were estimated more precisely by iterative algorithms. Advanced classifiers exploiting timing information were proposed (Cairns and Hansen, 1994; Womack and

* Corresponding author. Tel./fax: +30 2310 998225.
E-mail address: costas@aiia.csd.auth.gr (C. Kotropoulos).

0167-6393/$ - see front matter
doi:10.1016/j.specom.2006.04.003

Hansen, 1996; Polzin and Waibel, 1998). Nowadays, research is focused on finding powerful combinations of classifiers that advance the classification efficiency in real-life applications. The wide use of telecommunication services and multimedia devices paves also the way for new applications. For example, in the projects "Prosody for dialogue systems" and "SmartKom", ticket reservation systems are developed that employ automatic speech recognition being able to recognize the annoyance or frustration of a user and change their response accordingly (Ang et al., 2002; Schiel et al., 2002). Similar scenarios are also presented for call center applications (Petrushin, 1999; Lee and Narayanan, 2005). Emotional speech recognition can be employed by therapists as a diagnostic tool in medicine (France et al., 2000). In psychology, emotional speech recognition methods can cope with the bulk of enormous speech data in real-time extracting the speech characteristics that convey emotion and attitude in a systematic manner (Mozziconacci and Hermes, 2000).

In the future, the emotional speech research will primarily be benefited by the on-going availability of large-scale emotional speech data collections, and will focus on the improvement of theoretical models for speech production (Flanagan, 1972) or models related to the vocal communication of emotion (Scherer, 2003). Indeed, on the one hand, large data collections which include a variety of speaker utterances under several emotional states are necessary in order to faithfully assess the performance of emotional speech recognition algorithms. The already available data collections consist only of few utterances, and therefore it is difficult to demonstrate reliable emotion recognition results. The data collections listed in Section 2 provide initiatives to set up more relaxed and close to real-life specifications for recording large-scale emotional speech data collections that are complementary to the already existing resources. On the other hand, theoretical models of speech production and vocal communication of emotion will provide the necessary background for a systematic study and will deploy more accurate emotional cues through time. In the following, the contributions of the paper are identified and its outline is given.

1.1. Contributions of the paper

Several reviews on emotional speech analysis have already appeared (Van Bezooijen, 1984; Scherer et al., 1991; Cowie et al., 2001, 2003; Scherer, 2003; Douglas-Cowie et al., 2003). However, as the research towards understanding human emotions increasingly attracts the attention of the research community, the short list of 19 data collections appeared in (Douglas-Cowie et al., 2003) does not adequately cover the topic. In this tutorial, 64 data collections are reviewed. Furthermore, an up-to-date literature survey is provided, complementing the previous studies in (Van Bezooijen, 1984; Scherer et al., 1991; Cowie et al., 2001). Finally, the paper is focused on describing the feature extraction methods and the emotion classification techniques, topics that have not been treated in (Scherer, 2003; Pantic and Rothkrantz, 2003).

1.2. Outline

In Section 2, a corpus of 64 data collections is reviewed putting emphasis on the data collection procedures, the kind of speech (natural, simulated, or elicited), the content, and other physiological signals that may accompany the emotional speech. In Section 3, short-term features (i.e. features that are extracted on speech frame basis) that are related to the emotional content of speech are discussed. In addition to short-term features, their contours are of fundamental importance for emotional speech recognition. The emotions affect the contour characteristics, such as statistics and trends as is summarized in Section 4. Emotion classification techniques that exploit timing information and other techniques that ignore it are surveyed in Section 5. Therefore, Sections 3 and 4 aim at describing the appropriate features to be used with the emotional classification techniques reviewed in Section 5. Finally, Section 6 concludes the tutorial by indicating future research directions.

2. Data collections

A record of emotional speech data collections is undoubtedly useful for researchers interested in emotional speech analysis. An overview of 64 emotional speech data collections is presented in Table 1. For each data collection additional information is also described such as the speech language, the number and the profession of the subjects, other physiological signals possibly recorded simultaneously with speech, the data collection purpose (emotional speech recognition, expressive synthesis), the emotional states recorded, and the kind of the emotions (natural, simulated, elicited).

Table 1
Emotional speech data collections (in alphabetical ordering of the related references)

Reference	Language	Subjects	Other signals	Purpose	Emotions	Kind
Abelin and Allwood (2000)	Swedish	1 Native	–	Recognition	Ar, Fr, Jy, Sd, Se, Dt, Dom, Sy	Simulated
Alpert et al. (2001)	English	22 Patients, 19 healthy	–	Recognition	Dn, Nl	Natural
Alter et al. (2000)	German	1 Female	EEG	Recognition	Ar, Hs, Nl	Simulated
Ambrus (2000), Interface	English, Slovenian	8 Actors	LG	Synthesis	Ar, Dt, Fr, Nl, Se	Simulated
Amir et al. (2000)	Hebrew	40 Students	LG,M, G,H	Recognition	Ar, Dt, Fr, Jy, Sd	Natural
Ang et al. (2002)	English	Many	–	Recognition	An, At, Nl, Fd, Td	Natural
Banse and Scherer (1996)	German	12 Actors	V	Recognition	H/C Ar, Hs, Sd,...	Simulated
Batliner et al. (2004)	German, English	51 Children	–	Recognition	Ar, Bm, Jy, Se	Elicited
Bulut et al. (2002)	English	1 Actress	–	Synthesis	Ar, Hs, Nl, Sd	Simulated
Burkhardt and Sendlmeier (2000)	German	10 Actors	V, LG	Synthesis	Ar, Fr, Jy, Nl, Sd, Bm, Dt	Simulated
Caldognetto et al. (2004)	Italian	1 Native	V, IR	Synthesis	Ar, Dt, Fr, Jy, Sd, Se	Simulated
Choukri (2003), Groningen	Dutch	238 Native	LG	Recognition	Unknown	Simulated
Chuang and Wu (2002)	Chinese	2 Actors	–	Recognition	Ar, Ay, Hs, Fr, Se, Sd	Simulated
Clavel et al. (2004)	English	18 From TV	–	Recognition	Nl, levels of Fr	Simulated
Cole (2005), Kids' Speech	English	780 Children	V	Recognition, Synthesis	Unknown	Natural
Cowie and Douglas-Cowie (1996), Belfast Structured	English	40 Native	–	Recognition	Ar, Fr, Hs, Nl, Sd	Natural
Douglas-Cowie et al. (2003), Belfast Natural	English	125 From TV	V	Recognition	Various	Semi-natural
Edgington (1997)	English	1 Actor	LG	Synthesis	Ar, Bm, Fr, Hs, Nl, Sd	Simulated
Engberg and Hansen (1996), DES	Danish	4 Actors	–	Synthesis	Ar, Hs, Nl, Sd, Se	Simulated
Fernandez and Picard (2003)	English	4 Drivers	–	Recognition	Nl, Ss	Natural
Fischer (1999), Verbmobil	German	58 Native	–	Recognition	Ar, Dn, Nl	Natural
France et al. (2000)	English	70 Patients, 40 healthy	–	Recognition	Dn, Nl	Natural
Gonzalez (1999)	English, Spanish	Unknown	–	Recognition	Dn, Nl	Elicited
Hansen (1996), SUSAS	English	32 Various	–	Recognition	Ar, Ld eff., Ss, Tl	Natural, simulated
Hansen (1996), SUSC-0	English	18 Non-native	H,BP,R	Recognition	Nl, Ss	A-stress
Hansen (1996), SUSC-1	English	20 Native	–	Recognition	Nl, Ss	P-stress
Hansen (1996), DLP	English	15 Native	–	Recognition	Nl, Ss	C-stress
Hansen (1996), DCIEM	English	Unknown	–	Recognition	Nl, Sleep deprive	Elicited
Heuft et al. (1996)	German	3 Native	–	Synthesis	Ar, Fr, Jy, Sd,...	Simulated, elicited
Iida et al. (2000), ESC	Japanese	2 Native	–	Synthesis	Ar, Jy, Sd	Simulated
Iriondo et al. (2000)	Spanish	8 Actors	–	Synthesis	Fr, Jy, Sd, Se,...	Simulated

(*continued on next page*)

Table 1 (*continued*)

Reference	Language	Subjects	Other signals	Purpose	Emotions	Kind
Kawanami et al. (2003)	Japanese	2 Actors	–	Synthesis	Ar, Hs, Nl, Sd	Simulated
Lee and Narayanan (2005)	English	Unknown	–	Recognition	Negative–positive	Natural
Liberman (2005), Emotional Prosody	English	Actors	–	Unknown	Anxty, H/C Ar, Hs, Nl, Pc, Sd, Se,...	Simulated
Linnankoski et al. (2005)	English	13 Native	–	Recognition	An, Ar, Fr, Sd,...	Elicited
Lloyd (1999)	English	1 Native	–	Recognition	Phonological stress	Simulated
Makarova and Petrushin (2002), RUSSLANA	Russian	61 Native	–	Recognition	Ar, Hs, Se, Sd, Fr, Nl	Simulated
Martins et al. (1998), BDFALA	Portuguese	10 Native	–	Recognition	Ar, Dt, Hs, Iy	Simulated
McMahon et al. (2003), ORESTEIA	English	29 Native	–	Recognition	Ae, Sk, Ss	Elicited
Montanari et al. (2004)	English	15 Children	V	Recognition	Unknown	Natural
Montero et al. (1999), SES	Spanish	1 Actor	–	Synthesis	Ar, Dt, Hs, Sd	Simulated
Mozziconacci and Hermes (1997)	Dutch	3 Native	–	Recognition	Ar, Bm, Fr, Jy, Iy, Nl, Sd	Simulated
Niimi et al. (2001)	Japanese	1 Male	–	Synthesis	Ar, Jy, Sd	Simulated
Nordstrand et al. (2004)	Swedish	1 Native	V, IR	Synthesis	Hs, Nl	Simulated
Nwe et al. (2003)	Chinese	12 Native	–	Recognition	Ar, Fr, Dt, Jy,...	Simulated
Pereira (2000)	English	2 Actors	–	Recognition	H/C Ar, Hs, Nl, Sd	Simulated
Petrushin (1999)	English	30 Native	–	Recognition	Ar, Fr, Hs, Nl, Sd	Simulated, Natural
Polzin and Waibel (2000)	English	Unknown	–	Recognition	Ar, Fr, Nl, Sd	Simulated
Polzin and Waibel (1998)	English	5 Drama students	LG	Recognition	Ar, Fr, Hs, Nl, Sd	Simulated
Rahurkar and Hansen (2002), SOQ	English	6 Soldiers	H, R, BP, BL	Recognition	5 Stress levels	Natural
Scherer (2000b), Lost Luggage	Various	109 Passengers	V	Recognition	Ar, Hr, Ie, Sd, Ss	Natural
Scherer (2000a)	German	4 Actors	–	Ecological	Ar, Dt, Fr, Jy, Sd	Simulated
Scherer et al. (2002)	English, German	100 Native	–	Recognition	2 Tl, 2 Ss	Natural
Schiel et al. (2002), SmartKom	German	45 Native	V	Recognition	Ar, Dfn, Nl	Natural
Schröder and Grice (2003)	German	1 Male	–	Synthesis	Soft, modal, loud	Simulated
Schröder (2000)	German	6 Native	–	Recognition	Ar, Bm, Dt, Wy,...	Simulated
Slaney and McRoberts (2003), Babyears	English	12 Native	–	Recognition	Al, An, Pn	Natural
Stibbard (2000), Leeds	English	Unknown	–	Recognition	Wide range	Natural, elicited
Tato (2002), AIBO	German	14 Native	–	Synthesis	Ar, Bm, Hs, Nl, Sd	Elicited
Tolkmitt and Scherer (1986)	German	60 Native	–	Recognition	Cognitive Ss	Elicited
Wendt and Scheich (2002), Magdeburger	German	2 Actors	–	Recognition	Ar, Dt, Fr, Hs, Sd	Simulated
Yildirim et al. (2004)	English	1 Actress	–	Recognition	Ar, Hs, Nl, Sd	Simulated
Yu et al. (2001)	Chinese	Native from TV	–	Recognition	Ar, Hs, Nl, Sd	Simulated
Yuan (2002)	Chinese	9 Native	–	Recognition	Ar, Fr, Jy, Nl, Sd	Elicited

Abbreviations for emotions: The emotion categories are abbreviated by a combination of the first and last letters of their name. At: Amusement, Ay: Antipathy, Ar: Anger, Ae: Annoyance, Al: Approval, An: Attention, Anxty: Anxiety, Bm: Boredom, Dfn: Dissatisfaction, Dom: Dominance, Dn: Depression, Dt: Disgust, Fd: Frustrated, Fr: Fear, Hs: Happiness, Ie: Indifference, Iy: Irony, Jy: Joy, Nl: Neutral, Pc: Panic, Pn: Prohibition, Se: Surprise, Sd: Sadness, Ss: Stress, Sy: Shyness, Sk: Shock, Td: Tiredness, Tl: Task load stress, Wy: Worry. Ellipses denote that additional emotions were recorded.

Abbreviations for other signals: BP: Blood pressure, BL: Blood examination, EEG: Electroencephalogram, G: Galvanic skin response, H: Heart beat rate, IR: Infrared Camera, LG: Laryngograph, M: Myogram of the face, R: Respiration, V: Video.

Other abbreviations: H/C: Hot/cold, Ld eff.: Lombard effect, A-stress, P-stress, C-stress: Actual, Physical, and Cognitive stress, respectively, Sim.: Simulated, Elic.:Elicited, N/A: Not available.

From Table 1, it is evident that the research on emotional speech recognition is limited to certain emotions. The majority of emotional speech data collections encompasses five or six emotions, although the emotion categories are much more in real life. For example, many words "with emotional connotation", originally found in the semantic Atlas of Emotional Concepts, are enlisted in (Cowie and Cornelius, 2003). In the early 1970s, the pallet theory was proposed by Anscombe and Geach in an attempt to describe all emotions as a mixture of some primary emotions like what exactly happens with colors (Anscombe and Geach, 1970). This idea has been rejected and the term "*basic emotions*" is now widely used without implying that such emotions can be mixed to produce others (Eckman, 1992). It is commonly agreed that the basic emotions are more primitive and universal than the others. Eckman proposed the following basic emotions: anger, fear, sadness, sensory pleasure, amusement, satisfaction, contentment, excitement, disgust, contempt, pride, shame, guilt, embarrassment, and relief. Non-basic emotions are called "higher-level" emotions (Buck, 1999) and they are rarely represented in emotional speech data collections.

Three kinds of speech are observed. Natural speech is simply spontaneous speech where all emotions are real. Simulated or acted speech is speech expressed in a professionally deliberated manner. Finally, elicited speech is speech in which the emotions are induced. The elicited speech is neither neutral nor simulated. For example, portrayals of non-professionals while imitating a professional produce elicited speech, which can also be an acceptable solution when an adequate number of professionals is not available (Nakatsu et al., 1999). Acted speech from professionals is the most reliable for emotional speech recognition because professionals can deliver speech colored by emotions that possess a *high arousal*, i.e. emotions with a great amplitude or strength.

Additional synchronous physiological signals such as sweat indication, heart beat rate, blood pressure, and respiration could be recorded during the experiments. They provide a ground truth for the degree of subjects' arousal or stress (Rahurkar and Hansen, 2002; Picard et al., 2001). There is a direct evidence that the aforementioned signals are related more to the arousal information of speech than to the valence of the emotion, i.e. the positive or negative character of the emotions (Wagner et al., 2005). As regards other physiological signals, such as EEG or signals derived from blood analysis, no sufficient and reliable results have been reported yet.

The recording scenarios employed in data collections are presumably useful for repeating or augmenting the experiments. Material from radio or television is always available (Douglas-Cowie et al., 2003). However, such material raises copyright issues and impedes the data collection distribution. An alternative is speech from interviews with specialists, such as psychologists and scientists specialized in phonetics (Douglas-Cowie et al., 2003). Furthermore, speech from real-life situations such as oral interviews of employees when they are examined for promotion can be also used (Rahurkar and Hansen, 2002). Parents talking to infants, when they try to keep them away from dangerous objects can be another real-life example (Slaney and McRoberts, 2003). Interviews between a doctor and a patient before and after medication was used in (France et al., 2000). Speech can be recorded while the subject faces a machine, e.g. during telephone calls to automatic speech recognition (ASR) call centers (Lee and Narayanan, 2005), or when the subjects are talking to fake-ASR machines, which are operated by a human (wizard-of-OZ method, WOZ) (Fischer, 1999). Giving commands to a robot is another idea explored (Batliner et al., 2004). Speech can be also recorded during imposed stressed situations. For example when the subject adds numbers while driving a car at various speeds (Fernandez and Picard, 2003), or when the subject reads distant car plates on a big computer screen (Steeneken and Hansen, 1999). Finally, subjects' readings of emotionally neutral sentences located between emotionally biased ones can be another manner of recording emotional speech.

3. Estimation of short-term acoustic features

Methods for estimating short-term acoustic features that are frequently used in emotion recognition are described hereafter. Short-term features are estimated on a frame basis

$$f_s(n;m) = s(n)w(m-n), \quad (1)$$

where $s(n)$ is the speech signal and $w(m-n)$ is a window of length N_w ending at sample m (Deller et al., 2000). Most of the methods stem from the front-end signal processing employed in speech recognition and coding. However, the discussion is focused on acoustic features that are useful for emotion recognition. The outline of this section is as follows. Methods for estimating the fundamental phonation or pitch are discussed in Section 3.1. In Section 3.2 features based on a non-linear model of speech production are addressed. Vocal tract features related to emotional speech are described in Section 3.3. Finally, a method to estimate speech energy is presented in Section 3.4.

3.1. Pitch

The *pitch signal*, also known as the glottal waveform, has information about emotion, because it depends on the tension of the vocal folds and the subglottal air pressure. The pitch signal is produced from the vibration of the vocal folds. Two features related to the pitch signal are widely used, namely the pitch frequency and the glottal air velocity at the vocal fold opening time instant. The time elapsed between two successive vocal fold openings is called *pitch period T*, while the vibration rate of the vocal folds is the *fundamental frequency of the phonation* F_0 or *pitch frequency*. The glottal volume velocity denotes the air velocity through glottis during the vocal fold vibration. High velocity indicates a music like speech like joy or surprise. Low velocity is in harsher styles such as anger or disgust (Nogueiras et al., 2001). Many algorithms for estimating the pitch signal exist (Hess, 1992). Two algorithms will be discussed here. The first pitch estimation algorithm is based on the autocorrelation and it is the most frequent one. The second algorithm is based on a wavelet transform. It has been designed for stressed speech.

A widely spread method for extracting pitch is based on the *autocorrelation of center-clipped* frames (Sondhi, 1968). The signal is low filtered at 900 Hz and then it is segmented to short-time frames of speech $f_s(n;m)$. The clipping, which is a non-linear procedure that prevents the first formant interfering with the pitch, is applied to each frame $f_s(n;m)$ yielding

$$\hat{f}_s(n;m) = \begin{cases} f_s(n;m) - C_{thr} & \text{if } |f_s(n;m)| > C_{thr}, \\ 0 & \text{if } |f_s(n;m)| < C_{thr}, \end{cases} \tag{2}$$

where C_{thr} is set at the 30% of the maximum value of $f_s(n;m)$. After calculating the short-term autocorrelation

$$r_s(\eta;m) = \frac{1}{N_w} \sum_{n=m-N_w+1}^{m} \hat{f}_s(n;m)\hat{f}_s(n-\eta;m), \tag{3}$$

where η is the lag, the pitch frequency of the frame ending at m can be estimated by

$$\widehat{F}_0(m) = \frac{F_s}{N_w} \operatorname*{argmax}_{\eta} \{|r(\eta;m)|\}_{\eta=N_w(F_l/F_s)}^{\eta=N_w(F_h/F_s)}, \tag{4}$$

where F_s is the sampling frequency, and F_l, F_h are the lowest and highest perceived pitch frequencies by humans, respectively. Typical values of the aforementioned parameters are $F_s = 8000$ Hz, $F_l = 50$ Hz, and $F_h = 500$ Hz. The maximum value of the autocorrelation $\left(\max\{|r(\eta;m)|\}_{\eta=N_w(F_l/F_s)}^{\eta=N_w(F_h/F_s)}\right)$ is used as a measurement of the glottal velocity volume during the vocal fold opening (Nogueiras et al., 2001).

The autocorrelation method for pitch estimation was used with low error in emotion classification (Tolkmitt and Scherer, 1986; Iida et al., 2003). However, it is argued that this method of extracting pitch is affected by the interference of the first formant in the pitch frequency, no matter which the parameters of the center clipping are (Tolkmitt and Scherer, 1986). The clipping of small signal values may not remove the effect of the non-linear propagation of the air through the vocal tract which is an indication of the abnormal spectral characteristics during emotion.

The second method for estimating the pitch uses the wavelet transform (Cairns and Hansen, 1994). It is a derivation of the method described in (Kadambe and Boudreaux-Bartels, 1992). The pitch period extraction is based on a two pass dyadic wavelet transform over the signal. Let b denote a time index, 2^j be a scaling parameter, $s(n)$ be the sampled speech signal, and $\phi(n)$ be a cubic spline wavelet generated with the method in (Mallat and Zhong, 1989). The dyadic wavelet transform is defined by

$$D_yWT(b,2^j) = \frac{1}{2^j} \sum_{n=-\infty}^{n=+\infty} s(n)\phi\left(\frac{n-b}{2^j}\right). \tag{5}$$

It represents a convolution of the time-reversed wavelet with the speech signal. This procedure is repeated for three wavelet scales. In the first pass, the result of the transform is windowed by a 16 ms rectangular window shifted with a rate of 8 ms. The pitch frequency is found by estimating the maxima of $D_yWT(b,2^j)$ across the three scales. Although the method tracks the pitch epochs for neutral speech, it skips epochs for stressed speech. For marking the speech epochs in stressed speech, a second pass of wavelets is invented. In the second pass, the same wavelet transform is applied only in the intervals between the first pass pitch periods found to have a pitch epoch greater than 150% of the median value of the pitch epochs measured during the first pass. The result of the second wavelet transform is windowed by a 8 ms window with a 4 ms skip rate to capture the sudden pitch epochs that occur often in stressed speech.

The pitch period and the glottal volume velocity at the time instant of vocal fold opening are not the only characteristics of the glottal waveform. The shape of the glottal waveform during a pitch period is also informative about the speech signal and probably has to do with the emotional coloring of the speech, a topic that has not been studied adequately yet.

3.2. Teager energy operator

Another useful feature for emotion recognition is the *number of harmonics* due to the non-linear air flow in the vocal tract that produces the speech signal. In the emotional state of anger or for stressed speech, the fast air flow causes vortices located near the false vocal folds providing additional excitation signals other than the pitch (Teager and Teager, 1990; Zhou et al., 2001). The additional excitation signals are apparent in the spectrum as harmonics and cross-harmonics. In the following, a procedure to calculate the number of harmonics in the speech signal is described.

Let us assume that a speech frame $f_s(n;m)$ has a single harmonic which can be considered as an AM–FM sinewave. In discrete time, the AM–FM sinewave $f_s(n;m)$ can be represented as (Quatieri, 2002)

$$f_s(n;m) = \alpha(n;m)\cos(\Phi(n;m)) = \alpha(n;m) \times \cos\left(\omega_c n + \omega_h \int_0^n q(k)\mathrm{d}k + \theta\right) \quad (6)$$

with *instantaneous amplitude* $\alpha(n;m)$ and *instantaneous frequency*

$$\omega_i(n;m) = \frac{\mathrm{d}\Phi(n;m)}{\mathrm{d}n} = \omega_c + \omega_h q(n), \quad |q(n)| \leqslant 1, \quad (7)$$

where ω_c is the *carrier* frequency, $\omega_h \in [0,\omega_c]$ is the *maximum frequency deviation*, and θ is a constant phase offset.

The *Teager energy operator* (TEO) (Teager and Teager, 1990)

$$\Psi[f_s(n;m)] = (f_s(n;m))^2 - f_s(n+1;m)f_s(n-1;m) \quad (8)$$

when applied to an AM–FM sinewave yields the squared product of the AM–FM components

$$\Psi[f_s(n;m)] = \alpha^2(n;m)\sin(\omega_i^2(n;m)). \quad (9)$$

The unknown parameters $\alpha(n;m)$ and $\omega_i(n;m)$ can be estimated approximately with

$$\omega_i(n;m) \approx \arcsin\left(\sqrt{\frac{\Psi[\Delta_2]}{4\Psi[f_s(n;m)]}}\right) \quad (10)$$

and

$$\alpha(n;m) \approx \frac{2\Psi[f_s(n;m)]}{\sqrt{\Psi[\Delta_2]}}, \quad (11)$$

where $\Delta_2 = f_s(n+1;m) - f_s(n-1;m)$. Let us assume that within a speech frame each harmonic has an almost constant instantaneous amplitude and constant instantaneous frequency. If the signal has a single harmonic, then from (9) it is deduced that the TEO profile is a constant number. Otherwise, if the signal has more than one harmonic then the TEO profile is a function of n.

Since it is certain that more than one harmonic exist in the spectrum, it is more convenient to break the bandwidth into 16 small bands, and study each band independently. The polynomial coefficients, which describe the TEO autocorrelation envelope area, can be used as features for classifying the speech into emotional states (Zhou et al., 2001). This method achieves a correct classification rate of 89% in classifying neutral versus stressed speech whereas MFCCs yield 67% in the same task.

Pitch frequency also affects the number of harmonics in the spectrum. Less harmonics are produced when the pitch frequency is high. More harmonics are expected when the pitch frequency is low. It seems that the harmonics from the additional excitation signals due to vortices are more intense than those caused by the pitch signal. The interaction of the two factors is a topic for further research. A method which can be used to alleviate the presence of harmonics due to the pitch frequency factor is to normalize the speech so that it has a constant pitch frequency (Cairns and Hansen, 1994).

3.3. Vocal tract features

The shape of the vocal tract is modified by the emotional states. Many features have been used to describe the shape of the vocal tract during emotional speech production. Such features include

- the formants which are a representation of the vocal tract resonances,

- the cross-section areas when the vocal tract is modeled as a series of concatenated lossless tubes (Flanagan, 1972),
- the coefficients derived from frequency transformations.

The formants are one of the quantitative characteristics of the vocal tract. In the frequency domain, the location of vocal tract resonances depends upon the shape and the physical dimensions of the vocal tract. Since the resonances tend to ''form'' the overall spectrum, speech scientists refer to them as formants. Each formant is characterized by its center frequency and its bandwidth. It has been found that subjects during stress or under depression do not articulate voiced sounds with the same effort as in the neutral emotional state (Tolkmitt and Scherer, 1986; France et al., 2000). The formants can be used to discriminate the improved articulated speech from the slackened one. The formant bandwidth during slackened articulated speech is gradual, whereas the formant bandwidth during improved articulated speech is narrow with steep flanks. Next, we describe methods to estimate formant frequencies and formant bandwidths.

A simple method to estimate the formants relies on *linear prediction analysis*. Let an M-order all-pole vocal tract model with *linear prediction coefficients* (LPCs) $\hat{a}(i)$ be

$$\widehat{\Theta}(z) = \frac{1}{1 - \sum_{i=1}^{M} \hat{a}(i) z^{-i}}. \tag{12}$$

The angles of $\widehat{\Theta}(z)$ poles which are further from the origin in the z-plane are indicators of the formant frequencies (Atal and Schroeder, 1967; Markel and Gray, 1976). When the distance of a pole from the origin is large then the bandwidth of the corresponding formant is narrow with steep flanks, whereas when a pole is close to the origin then the bandwidth of the corresponding formant is wide with gradual flanks. Experimental analysis has shown that the first and second formants are affected by the emotional states of speech more than the other formants (Tolkmitt and Scherer, 1986; France et al., 2000).

A problem faced with the LPCs in formant tracking procedure is the false identification of the formants. For example, during the emotional states of happiness and anger, the second formant (F_2) is confused with the first formant (F_1) and F_1 interferes with the pitch frequency (Yildirim et al., 2004). A formant tracking method which does not suffer from the aforementioned problems is proposed in (Cairns and Hansen, 1994), which was originally developed by Hanson et al. (1994). Hanson et al. (1994) found that an approximate estimate of a formant location, $\omega_i(n;m)$ calculated by (10), could be used to iteratively refine the formant center frequency via

$$f_c^{l+1}(m) = \frac{1}{2\pi N_w} \sum_{n=m-N_w+1}^{m} \omega_i(n;m), \tag{13}$$

where $f_c^{l+1}(m)$ is the formant center frequency during iteration $l+1$. If the distance between $f_c^{l+1}(m)$ and $f_c^{l}(m)$ is smaller than 10 Hz, then the method stops and f_c^{l+1} is the formant frequency estimate. In detail, $f_c^{1}(m)$ is estimated by the formant frequency estimation procedure that employs LPCs. The signal is filtered with a bandpass filter in order to isolate the band which includes the formant. Let $G_l(n)$ be the impulse response of a Gabor bandpass filter

$$G_l(n) = \exp[-(\beta n T)^2] \cos(2\pi f_c^l T n), \tag{14}$$

where f_c^l is the center frequency, β the bandwidth of the filter, and T is the sampling period. If $f_c^l < 1000$ Hz, then β equals to 800 Hz, otherwise $\beta = 1100$ Hz. The value of β is chosen small enough so as not to have more than one formant inside the bandwidth and large enough to capture the change of the instantaneous frequency. Then, f_c^{l+1} is estimated by (13). If the criterion $|f_c^{l+1} - f_c^l| < 10$ is satisfied, then the method stops, otherwise the frame is refiltered with the Gabor filter centered at f_c^{l+1}. The latter is re-estimated with (13) and the criterion is checked again. The method stops after a few iterations. However, it is reported that there are a few exceptions where the method does not converge. This could be a topic for further study.

The second feature is the cross-section areas of the vocal tract modeled by the multi-tube lossless model (Flanagan, 1972). Each tube is described by its cross-section area and its length. To a first approximation, one may assume that there is no loss of energy due to soft wall vibrations, heat conduction, and thermal viscosity. For a large number of tubes, the model becomes a realistic representation of the vocal tract, but it is not possible to be computed in real time. A model that can easily be computed is that of 10 cross-section areas of fixed length (Mrayati et al., 1988). The cross-section area near the glottis is indexed by A_1 and the others are following sequentially until the lips. The back vocal tract area A_2 can be used to discriminate the neutral speech from that by anger colored, as A_2 is greater in the former emotion than in the latter one (Womack and Hansen, 1996).

The third feature is the energy of certain frequency bands. There are many contradictions in identifying the best frequency band of the power spectrum in order to classify emotions. Many investigators put high significance on the low frequency bands, such as the 0–1.5 kHz band (Tolkmitt and Scherer, 1986; Banse and Scherer, 1996; France et al., 2000) whereas other suggest the opposite (Nwe et al., 2003). An explanation for both opinions is that stressed or colored by anger speech may be expressed with a low articulation effort, a fact which causes formant peak smoothing and spectral flatness as well as energy shifting from low to high frequencies in the power spectrum. The *Mel-frequency cepstral coefficients* (MFCCs) (Davis and Mermelstein, 1980) provide a better representation of the signal than the frequency bands since they additionally exploit the human auditory frequency response. Nevertheless, the experimental results have demonstrated that the MFCCs achieve poor emotion classification results (Zhou et al., 2001; Nwe et al., 2003), which might be due to the textual dependency and the embedded pitch filtering during cepstral analysis (Davis and Mermelstein, 1980). Better features than MFCCs for emotion classification in practice are the log-frequency power coefficients (LFPCs) which include the pitch information (Nwe et al., 2003). The LFPCs are simply derived by filtering each short-time spectrum with 12 bandpass filters having bandwidths and center frequencies corresponding to the critical bands of the human ear (Rabiner and Juang, 1993).

3.4. Speech energy

The short-term speech energy can be exploited for emotion recognition, because it is related to the arousal level of emotions. The short-term energy of the speech frame ending at m is

$$E_s(m) = \frac{1}{N_w} \sum_{n=m-N_w+1}^{m} |f_s(n;m)|^2. \tag{15}$$

4. Cues to emotion

In this section, we review how the contour of selected short-term acoustic features is affected by the emotional states of anger, disgust, fear, joy, and sadness. A short-term feature contour is formed by assigning the feature value computed on a frame basis to all samples belonging to the frame. For example, the energy contour is given by

$$e(n) = E_s(m), \quad n = m - N_w + 1, \ldots, m. \tag{16}$$

The contour trends (i.e. its plateaux, its rising or falling slopes) is a valuable feature for emotion recognition, because they describe the temporal characteristics of an emotion. The survey is limited to those acoustic features for which at least two references are found in the literature (Van Bezooijen, 1984; Cowie and Douglas-Cowie, 1996; Pantic and Rothkrantz, 2003; Gonzalez, 1999; Heuft et al., 1996; Iida et al., 2000; Iriondo et al., 2000; Montero et al., 1999; Mozziconacci and Hermes, 2000; Murray and Arnott, 1996; Pollerman and Archinard, 2002; Scherer, 2003; Ververidis and Kotropoulos, 2004; Yuan, 2002). The following statistics are measured for the extracted features:

- mean, range, variance, and the pitch contour trends;
- mean and range of the intensity contour;
- rate of speech and transmission duration between utterances.

The speech rate is calculated as the inverse duration of the voiced part of speech determined by the presence of pitch pulses (Dellaert et al., 1996; Banse and Scherer, 1996) or it can be found by the rate of syllabic units. The speech signal can be segmented into syllabic units using the maxima and the minima of energy contour (Mermelstein, 1975).

In Table 2, the behavior of the most studied acoustic features for the five emotional states under consideration is outlined. Anger is the emotion of the highest energy and pitch level. Angry males show higher levels of energy than angry females. It is found that males express anger with a slow speech rate as opposed to females who employ a fast speech rate under similar circumstances (Heuft et al., 1996; Iida et al., 2000). Disgust is expressed with a low mean pitch level, a low intensity level, and a slower speech rate than the neutral state does. The emotional state of fear is correlated with a high pitch level and a raised intensity level. The majority of research outcomes reports a wide pitch range. The pitch contour has falling slopes and sometimes plateaux appear. The lapse of time between speech segments is shorter than that in the neutral state. Low levels of the mean intensity and mean pitch

Table 2
Summary of the effects of several emotion states on selected acoustic features

	Pitch				Intensity		Timing	
	Mean	Range	Variance	Contour	Mean	Range	Speech rate	Transmission duration
Anger	$>>$	$>$	$>>$		$>>_{M}, >_{F}$	$>$	$<_{M}, >_{F}$	$<$
Disgust	$<$	$>_{M}, <_{F}$			$<$		$\ll_{M}, <_{F}$	
Fear	$>>$	$>$		$\nearrow$	$=>$			$<$
Joy	$>$	$>$	$>$	$\searrow$	$>$	$>$		$<$
Sadness	$<$	$<$	$<$	$\nearrow$	$<$	$<$	$>_{M}, <_{F}$	$>$

Explanation of symbols: $>$: increases, $<$: decreases, $=$: no change from neutral, $\nearrow$: inclines, $\searrow$: declines. Double symbols indicate a change of increased predicted strength. The subscripts refer to gender information: M stands for males and F stands for females.

are measured when the subjects express sadness. The speech rate under similar circumstances is generally slower than that in the neutral state. The pitch contour trend is a valuable parameter, because it separates fear from joy. Fear resembles sadness having an almost downwards slope in the pitch contour, whereas joy exhibits a rising slope. The speech rate varies within each emotion. An interesting observation is that males speak faster when they are sad than when they are angry or disgusted.

The trends of prosody contours include discriminatory information about emotions. However, very few the efforts to describe the shape of feature contours in a systematic manner can be found in the literature. In (Leinonen et al., 1997; Linnankoski et al., 2005), several statistics are estimated on the syllables of the word 'Sarah'. However, there is no consensus if the results obtained from a word are universal due to textual dependency. Another option is to estimate feature statistics on the *rising* or *falling slopes* of contours as well as at their plateaux at *minima/maxima* (McGilloway et al., 2000; Ververidis and Kotropoulos, 2004; Bänziger and Scherer, 2005). Statistics such as the mean and the variance are rather rudimentary. An alternative is to transcribe the contour into discrete elements, i.e. a sequence of symbols that provide information about the tendency of a contour on a short-time basis. Such elements can be provided by the *ToBI* (Tones and Breaks Indices) system (Silverman et al., 1992). For example, the pitch contour is transcribed into a sequence of binary elements L, H, where L stands for low and H stands for high values, respectively. There is evidence that some sequences of L and H elements provide information about emotions (Stibbard, 2000). A similar investigation for 10 elements that describe the duration and the inclination of rising and falling slopes of pitch contour also exists (Mozziconacci and Hermes, 1997). Classifiers based on discrete elements have not been studied yet. In the following section, several techniques for emotion classification are described.

5. Emotion classification techniques

The output of emotion classification techniques is a prediction value (label) about the emotional state of an utterance. An utterance u_{ξ} is a speech segment corresponding to a word or a phrase. Let u_{ξ}, $\xi \in \{1, 2, \ldots, \Xi\}$ be an utterance of the data collection. In order to evaluate the performance of a classification technique, the cross-validation method is used. According to this method, the utterances of the whole data collection are divided into the design set $\mathscr{D}_s$ containing $N_{\mathscr{D}_s}$ utterances and the test set $\mathscr{T}_s$ comprised of $N_{\mathscr{T}_s}$ utterances. The classifiers are trained using the design set and the classification error is estimated on the test set. The design and the test set are chosen randomly. This procedure is repeated for several times defined by the user and the estimated classification error is the average classification error over all repetitions (Efron and Tibshirani, 1993).

The classification techniques can be divided into two categories, namely those employing

- prosody contours, i.e. sequences of short-time prosody features;
- statistics of prosody contours, like the mean, the variance, etc. or the contour trends.

The aforementioned categories will be reviewed independently in this section.

5.1. Classification techniques that employ prosody contours

The emotion classification techniques that employ prosody contours exploit the temporal information of speech, and therefore could be useful for speech recognition. Three emotion classification techniques were found in the literature, namely a technique based on artificial neural networks (ANNs) (Womack and Hansen, 1996), the *multi-channel hidden Markov Model* (Womack and Hansen, 1999), and the *mixture of hidden Markov models* (Fernandez and Picard, 2003).

In the first classification technique, the short-time features are used as an input to an ANN in order to classify utterances into emotional states (Womack and Hansen, 1996). The algorithm is depicted in Fig. 1. The utterance u_ξ is partitioned into Q bins containing K frames each. Q varies according to the utterance length, whereas K is a constant number. Let $x_{\xi q}$ denote a bin of u_ξ, where $q \in \{1, 2, \ldots, Q\}$. $x_{\xi q}$ is classified automatically to a phoneme group, such as fricatives (FR), vowels (VL), semi-vowels (SV), etc. by means of hidden Markov Models (HMMs) (Pellom and Hansen, 1996). Let Θ_λ denote the λth phoneme group, where $\lambda = 1, 2, \ldots, \Lambda$. From each frame $t = 1, 2, \ldots, K$ of the bin $x_{\xi q}$, D features related to the emotional state of speech are extracted. Let $y_{\xi qtd}$ be the dth feature of the tth frame for the bin $x_{\xi q}$, where $d \in \{1, 2, \ldots, D\}$. The $K \times D$ matrix of feature values is rearranged to a vector of length KD, by lexicographic ordering of the rows of the $K \times D$ matrix. This feature vector of KD feature values extracted from the bin $x_{\xi q}$ is input to the ANN described in Section 5.2. Let Ω_c be an emotional state, where $c \in \{1, 2, \ldots, C\}$. An ANN is trained on the cth emotional state of the λth phoneme group. The output node of the ANN denotes the likelihood of $x_{\xi q}$ given the emotional state Ω_c and the phoneme group Θ_λ. The likelihood of an utterance u_ξ given the emotional state Ω_c is the sum of the likelihoods for all $x_{\xi q} \in u_\xi$ given Ω_c and Θ_λ

$$P(u_\xi|\Omega_c) = \sum_{q=1}^{Q} \sum_{\lambda=1}^{\Lambda} P(x_{\xi q}|\Omega_c, \Theta_\lambda) P(\Theta_\lambda). \quad (17)$$

The aforementioned technique achieves a correct classification rate of 91% for 10 stress categories using vocal tract cross-section areas (Womack and Hansen, 1996). An issue for further study is the evolution of the emotional cues through time. Such a study can be accomplished through a new classifier which employs as input the output of each ANN.

The second emotion classification technique is called *multi-channel hidden Markov model* (Womack and Hansen, 1999). Let s_i, $i = 1, 2, \ldots, V$ be a sequence of states of a single-channel HMM. By using a single-channel HMM, a classification system can be described at any time as being in one of V distinct states that correspond to phonemes, as is presented in Fig. 2(a) (Rabiner and Juang, 1993). The multi-channel HMM combines the benefits of emotional speech classification with a traditional single-channel HMM for speech recognition. For example a C-channel HMM could be formulated to model speech from C emotional states with one dimension allocated for each emotional state, as is depicted in Fig. 2(b). In detail, the multi-channel

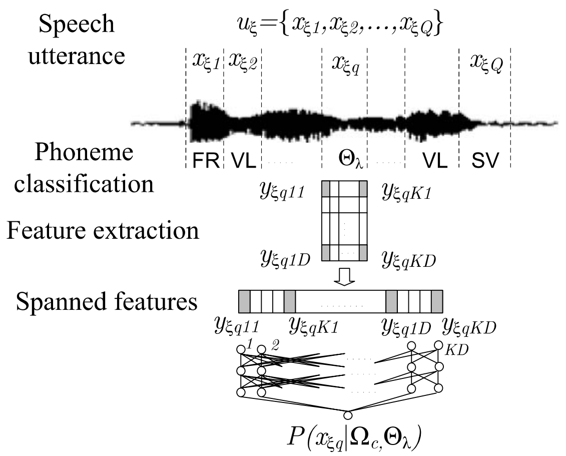

Fig. 1. An emotion classification technique that employs HMMs for phoneme classification and ANNs for emotion classification.

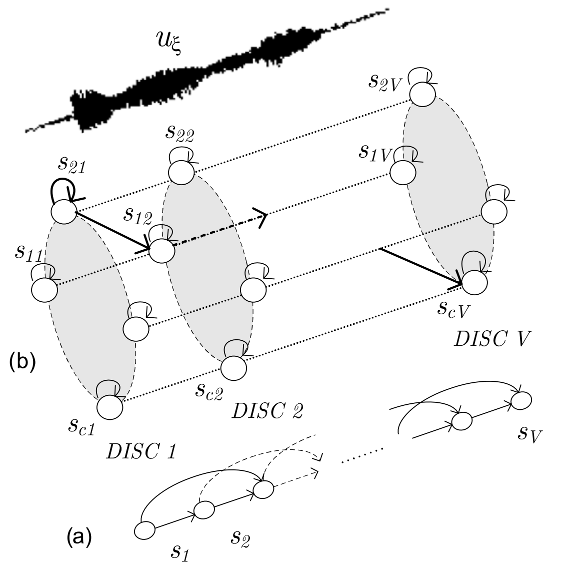

Fig. 2. Two structures of HMMs that can be used for emotion recognition: (a) a single-channel HMM and (b) a multi-channel HMM.

HMM consists of states s_{cv}, $v = 1, 2, \ldots, V$, $c = 1, 2, \ldots, C$. The states s_{cv}, $c = 1, 2, \ldots, C$ form a *disc*. Transitions are allowed from left to right as in a single-channel HMM, across emotional states within the same disc, and across emotional states in the next disc. It offers the additional benefit of a subphoneme speech model at the emotional state level instead of the phoneme level. The overall flexibility of the multi-channel HMM is improved by allowing a combined model where the integrity of each dimension is preserved (Womack and Hansen, 1999). In addition to a C mixture single-channel HMM it offers separate state transition probabilities.

The training phase of the multi-channel HMM consists of two steps. The first step requires training of each single-channel HMM to an emotional state, and the second step combines the emotion-dependent single-channel HMMs into a multi-channel HMM. In order to classify an utterance, a probability measurement is constructed. The likelihood of an utterance given an emotional state Ω_c is the ratio of the number of passes through states s_{cv}, $v = 1, 2, \ldots, V$ versus the total number of state transitions. The multi-channel HMM was used firstly for stress classification, and secondly for speech recognition on a data collection consisting of 35 words spoken in four stress styles. The correct stress classification rate achieved was 57.6% using MFCCs, which was almost equal to the stress classification rate of 58.6% achieved by the single-channel HMM using the same features. A reason for the aforementioned performance deterioration might be the small size of the data collection (Womack and Hansen, 1999). However, the multi-channel HMM achieved a correct speech classification rate of 94.4%, whereas the single-channel HMM achieved a rate of 78.7% in the same task. The great performance of the multi-channel HMM in speech recognition experiments might be an indication that the proposed model can be useful for stress classification in large data collections. A topic for further investigation would be to model the transitions across the disks with an additional HMM or an ANN (Bou-Ghazale and Hansen, 1998).

The third technique used for emotion classification is the so-called *mixture of HMMs* (Fernandez and Picard, 2003). The technique consists of two training stages. In the first stage, an unsupervised iterative clustering algorithm is used to discover M clusters in the feature space of the training data, where it is assumed that the data of each cluster are governed by a single underlying HMM. In the second stage, a number of HMMs are trained on the clusters. Each HMM is trained on the cth emotional state of the mth cluster, where $c = 1, 2, \ldots, C$ and $m = 1, 2, \ldots, M$. Both training stages and the classification of an utterance which belongs to the test set are described next.

In the first training stage, the utterances of the training set are divided into M clusters. Let $\Gamma^{(l)} = \{\gamma_1^{(l)}, \ldots, \gamma_m^{(l)}, \ldots, \gamma_M^{(l)}\}$ be the clusters at the lth iteration of the clustering algorithm, $\Delta^{(l)} = \{\delta_1^{(l)}, \ldots, \delta_m^{(l)}, \ldots, \delta_M^{(l)}\}$ be the HMM parameters for the cluster set $\Gamma^{(l)}$, $P(u_\xi|\delta_m^{(l)})$ be the likelihood of u_ξ given the cluster with HMM parameters $\delta_m^{(l)}$, and

$$P^{(l)} = \sum_{m=1}^{M} \sum_{u_\xi \in \gamma_m^{(l)}} \log P(u_\xi|\delta_m^{(l)}) \tag{18}$$

be the log-likelihood of all utterances during the lth iteration. The iterative clustering procedure is described in Fig. 3. In the second training stage, the utterances which have already been classified into a cluster γ_m are used to train C HMMs, where each HMM corresponds to an emotional state. Let $P(\delta_m|\Omega_c)$ be the ratio of the utterances that were assigned to cluster γ_m and belong to Ω_c over the number of the training utterances. In order to classify a test utterance u_ξ into an emotional state the Bayes classifier is used. The probability of an emotional state Ω_c given an utterance is

$$\begin{aligned} P(\Omega_c|u_\xi) &= \sum_{m=1}^{M} P(\Omega_c, \delta_m|u_\xi) \\ &= \sum_{m=1}^{M} P(u_\xi|\Omega_c, \delta_m) P(\delta_m|\Omega_c) P(\Omega_c), \end{aligned} \tag{19}$$

Step 1. Assign randomly the utterances to obtain the initial clusters $\Gamma^{(0)}$. Calculate $\Delta^{(0)}$ given $\Gamma^{(0)}$ using the Viterbi algorithm (Rabiner and Juang, 1993). Estimate $P^{(0)}$ using (18).
Step 2. Re-assign the utterances using $\Delta^{(0)}$ to the cluster with the highest likelihood in order to obtain $\Gamma^{(1)}$, i.e. assign u_ξ to $\gamma_{\hat{k}}^{(1)}$, where $\hat{k} = \arg\max_m P(u_\xi|\delta_m^{(0)})$. Calculate $\Delta^{(1)}$ from $\Gamma^{(1)}$. Estimate $P^{(1)}$.
Step 3. Re-assign the utterances using $\Delta^{(l-1)}$ to obtain $\Gamma^{(l)}$. Calculate $\Delta^{(l)}$ from $\Gamma^{(l)}$. Estimate $P^{(l)}$.
Step 4. If $|P^{(l)} - P^{(l-1)}| < \epsilon$ stop, where ϵ is a user-defined threshold. The procedure stops when (18) reaches a maximum. Otherwise, $l = l + 1$ and go to Step 3.

Fig. 3. A clustering procedure that it is based on HMMs.

where $P(u_\xi|\Omega_c,\delta_m)$ is the output of the HMM which was trained on the emotional state Ω_c of the cluster γ_m, and $P(\Omega_c)$ is the likelihood of each emotional state in the data collection. The correct classification rate achieved for four emotional states by the mixture of HMMs was 62% using energy contours in several frequency bands, whereas a single-channel HMM yields a smaller classification rate by 10% using the same features. A topic of future investigation might be the clustering algorithm described in Fig. 3. It is not clear what each cluster of utterances represents. Also, the convergence of the clustering procedure has not been investigated yet.

5.2. Classification techniques that employ statistics of prosody contours

Statistics of prosody contours have also been used as features for emotion classification techniques. The major drawback of such classification techniques is the loss of the timing information. In this section, the emotion classification techniques are separated into two classes, namely those that estimate the probability density function (pdf) of the features and those that discriminate emotional states without any estimation of the feature distributions for each emotional state. In Table 3, the literature related to discriminant classifiers applied to emotion recognition is summarized. First, the Bayes classifier when the class pdfs are modeled either as *Gaussians*, or *mixtures of Gaussians*, or estimated via *Parzen windows* is described. Next, we briefly discuss classifiers that do not employ any pdf modeling such as the *k-nearest neighbors*, the *support vector machines*, and the *artificial neural networks*.

The features used for emotion classification are statistics of the prosody contours such as the mean, the variance, etc. A full list of such features can be found in (Ververidis and Kotropoulos, 2004). Let $\boldsymbol{y}_\xi = (y_{\xi 1} y_{\xi 2} \cdots y_{\xi D})^{\mathrm{T}}$ be the *measurement vector* containing $y_{\xi d}$ statistics extracted from u_ξ, where $d = 1, 2, \ldots, D$ denotes the feature index.

According to the *Bayes classifier*, an utterance u_ξ is assigned to emotional state $\Omega_{\hat{c}}$, if

$$\hat{c} = \arg\max_{c=1}^{C}\{P(\boldsymbol{y}_\xi|\Omega_c)P(\Omega_c)\}, \tag{20}$$

where $P(\boldsymbol{y}|\Omega_c)$ is the pdf of $\boldsymbol{y}_\xi$ given the emotional state Ω_c, and $P(\Omega_c)$ is the *prior probability* of having the emotional state Ω_c. $P(\Omega_c)$ represents the knowledge we have about the emotional state of an utterance before the measurement vector of that utterance is available. Three methods for estimating $P(\boldsymbol{y}|\Omega_c)$ will be summarized, namely the single Gaussian model, the mixture of Gaussian densities model or Gaussian Mixture Model (GMM), and the estimation via Parzen windows.

Suppose that a measurement vector $\boldsymbol{y}_\xi$ coming from utterances that belong to Ω_c is distributed according to a single multi-variate *Gaussian* distribution:

$$P(\boldsymbol{y}|\Omega_c) = g(\boldsymbol{y};\boldsymbol{\mu}_c,\boldsymbol{\Sigma}_c) = \frac{\exp\left[-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu}_c)^{\mathrm{T}}\boldsymbol{\Sigma}_c^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_c)\right]}{(2\pi)^{D/2}|\det(\boldsymbol{\Sigma}_c)|^{1/2}}, \tag{21}$$

where $\boldsymbol{\mu}_c$, $\boldsymbol{\Sigma}_c$ are the mean vector and the covariance matrix, and det is the determinant of a matrix. The Bayes classifier, when the class conditional pdfs of the energy and pitch contour statistics are modeled by (21), achieves a correct classification rate of a

Table 3
Discriminant classifiers for emotion recognition

	Classifier	References
With pdf modeling	Bayes classifier using one Gaussian pdf	Dellaert et al. (1996), Schüller et al. (2004)
	Bayes classifier using one Gaussian pdf with linear discriminant analysis	France et al. (2000), Lee and Narayanan (2005)
	Bayes classifier using pdfs estimated by Parzen windows	Dellaert et al. (1996), Ververidis et al. (2004)
	Bayes classifier using a mixture of Gaussian pdfs	Slaney and McRoberts (2003), Schüller et al. (2004), Jiang and Cai (2004), Ververidis and Kotropoulos (2005)
Without pdf modeling	*K*-nearest neighbors	Dellaert et al. (1996), Petrushin (1999), Picard et al. (2001)
	Support vector machines	McGilloway et al. (2000), Fernandez and Picard (2003), Kwon et al. (2003)
	Artificial neural networks	Petrushin (1999), Tato (2002), Shi et al. (2003), Fernandez and Picard (2003), Schüller et al. (2004)

56% for four emotional states (Dellaert et al., 1996). The benefit of the Gaussian model is that it is estimated fast. Its drawback is that the assumption of Gaussian distributed features may not be true for real data. Linear discriminant analysis is a method to improve the classification rates achieved by the Bayes classifier, when each $P(\boldsymbol{y}|\Omega_c)$ is modeled as in (21).

In linear discriminant analysis the measurement space is transformed so that the separability between the emotional states is maximized. We will focus on the problem of two emotional states Ω_1 and Ω_2 to maintain simplicity. Let N_1 and N_2 be the number of utterances that belong to Ω_1 and Ω_2, respectively. The separability between the emotional states can be expressed by several criteria. One such criterion is the

$$J = \mathrm{tr}(\boldsymbol{S}_{\mathrm{w}}^{-1}\boldsymbol{S}_{\mathrm{b}}), \tag{22}$$

where $\boldsymbol{S}_{\mathrm{w}}$ is the within emotional states scatter matrix defined by

$$\boldsymbol{S}_{\mathrm{w}} = \frac{N_1}{N_{\mathrm{s}}}\boldsymbol{\Sigma}_1 + \frac{N_2}{N_{\mathrm{s}}}\boldsymbol{\Sigma}_2, \tag{23}$$

and $\boldsymbol{S}_{\mathrm{b}}$ is the between emotional states scatter matrix given by

$$\boldsymbol{S}_{\mathrm{b}} = \frac{N_1}{N_{\mathrm{s}}}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^{\mathrm{T}} + \frac{N_2}{N_{\mathrm{s}}}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_0)(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_0)^{\mathrm{T}}, \tag{24}$$

where $\boldsymbol{\mu}_0$ is the gross mean vector. A linear transformation $\boldsymbol{z} = A^{\mathrm{T}}\boldsymbol{y}$ of measurements from space Y to space Z which maximizes J is sought. The scatter matrices $\boldsymbol{S}_{\mathrm{b}Z}$ and $\boldsymbol{S}_{\mathrm{w}Z}$ in the Z-space are calculated from $\boldsymbol{S}_{\mathrm{b}}$ and $\boldsymbol{S}_{\mathrm{w}}$ in the Y-space by

$$\boldsymbol{S}_{\mathrm{b}Z} = A^{\mathrm{T}}\boldsymbol{S}_{\mathrm{b}Y}A, \qquad \boldsymbol{S}_{\mathrm{w}Z} = A^{\mathrm{T}}\boldsymbol{S}_{\mathrm{w}Y}A. \tag{25}$$

Thus, the problem of transformation is to find A which optimizes J in the Z-space. It can be shown that the optimum A is the matrix formed by the eigenvectors that correspond to the maximal eigenvalues of $\boldsymbol{S}_{\mathrm{w}Y}^{-1}\boldsymbol{S}_{\mathrm{b}Y}$. A linear discriminant classifier achieves a correct classification of 93% for two emotional classes using statistics of pitch and energy contours (Lee and Narayanan, 2005). Linear discrimination analysis has a disadvantage. The criterion in (22) may not be a good measure of emotional state separability when the pdf of each emotional state in the measurement space Y is not a Gaussian (21) (Fukunaga, 1990).

In the *GMM*, it is assumed that the measurement vectors $\boldsymbol{y}_\xi$ of an emotional state Ω_c are divided into clusters, and the measurement vectors in each cluster follow a Gaussian pdf. Let K_c be the number of clusters in the emotional state Ω_c. The complete pdf estimate is

$$P(\boldsymbol{y}|\Omega_c) = \sum_{k=1}^{K_c} g(\boldsymbol{y}; \boldsymbol{\mu}_{ck}, \boldsymbol{\Sigma}_{ck}) \tag{26}$$

which depends on the mean vector $\boldsymbol{\mu}_{ck}$, the covariance matrix $\boldsymbol{\Sigma}_{ck}$, and the mixing parameter π_{ck} $\left(\sum_{k=1}^{K_c}\pi_{ck} = 1,\ \pi_{ck} \geqslant 0\right)$ of the kth cluster in the cth emotional state. The parameters $\boldsymbol{\mu}_{ck}$,$\boldsymbol{\Sigma}_{ck}$,π_{ck} are calculated with the *expectation maximization* algorithm (EM) (Dempster et al., 1977), and K_c can be derived by the Akaike information criterion (Akaike, 1974). A correct classification rate of 75% for three emotional states is achieved by the Bayes classifier, when each $P(\boldsymbol{y}|\Omega_c)$ of pitch and energy contour statistics is modeled as a mixture of Gaussian densities (Slaney and McRoberts, 2003). The advantage of the Gaussian mixture modeling is that it might discover relationships between the clusters and the speakers. A disadvantage is that the EM converges to a local optimum.

By using *Parzen windows* an estimate of the $P(\boldsymbol{y}|\Omega_c)$ could also be obtained. It is certain that at $\boldsymbol{y}_\xi$ corresponding to $u_\xi \in \Omega_c$, $p(\boldsymbol{y}_\xi|\Omega_c) \neq 0$. Since an emotional state pdf is continuous over the measurement space, it is expected that $P(\boldsymbol{y}|\Omega_c)$ in the neighborhood of $\boldsymbol{y}_\xi$ should also be non-zero. The further we move away from $\boldsymbol{y}_\xi$, the less we can say about the $P(\boldsymbol{y}|\Omega_c)$. When using Parzen windows for class pdf estimation, the knowledge gained by the measurement vector $\boldsymbol{y}_\xi$ is represented by a function positioned at $\boldsymbol{y}_\xi$ and with an influence restricted to the neighborhood of $\boldsymbol{y}_\xi$. Such a function is called the *kernel* of the estimator. The kernel function $h(\cdot)$ can be any function from $\mathbb{R}^+ \to \mathbb{R}^+$ that admits a maximum at $\boldsymbol{y}_\xi$ and it is monotonically increasing as $\boldsymbol{y} \to \boldsymbol{y}_\xi$. Let $d(\boldsymbol{y},\boldsymbol{y}_\xi)$ be the Euclidean, Mahalanobis or any other appropriate distance measure. The pdf of an emotional state Ω_c is estimated by (van der Heijden et al., 2004)

$$P(\boldsymbol{y}|\Omega_c) = \frac{1}{N_c}\sum_{\boldsymbol{y}_\xi \in \Omega_c} h(d(\boldsymbol{y}, \boldsymbol{y}_\xi)). \tag{27}$$

A Bayes classifier achieves a correct classification rate of 53% for five emotional states, when each

$P(\boldsymbol{y}|\Omega_c)$ of pitch and energy contour statistics is estimated via Parzen windows (Ververidis et al., 2004). An advantage by estimating $P(\boldsymbol{y}|\Omega_c)$ via Parzen windows is that a prior knowledge about the conditional pdf of the measurement vectors is not required. The pdfs of the measurement vector for small data collections are hard to find. The execution time for modeling a conditional pdf by Parzen windows is relatively shorter than by a GMM estimated with the EM algorithm. A disadvantage is that the estimate of $P(\boldsymbol{y}|\Omega_c)$ has a great number of peaks that are not present in the real pdf.

A *support vector classifier* separates the emotional states with a maximal margin. The margin γ is defined by the width of the largest 'tube' not containing utterances that can be drawn around a decision boundary. The measurement vectors that define the boundaries of the margin are called *support vectors*. We shall confine ourselves to a two-class problem without any loss of generality. A support vector classifier was originally designed for a two-class problem, but it can be expanded to more classes.

Let us assume that a training set of utterances is denoted by $\{u_\xi\}_{\xi=1}^{N_{\mathscr{D}_s}} = \{(\boldsymbol{y}_\xi, l_\xi)\}_{\xi=1}^{N_{\mathscr{D}_s}}$, where $l_\xi \in \{-1, +1\}$ is the emotional state membership of each utterance. The classifier is a hyperplane

$$g(\boldsymbol{y}) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{y} + b, \tag{28}$$

where $\boldsymbol{w}$ is the gradient vector which is perpendicular to the hyperplane, and b is the offset of the hyperplane from the origin. It can be shown that the margin is inversely proportional to $\|\boldsymbol{w}\|^2/2$. The quantity $l_\xi g(\boldsymbol{y}_\xi)$ can be used to indicate to which side of the hyperplane the utterance belongs to. $l_\xi g(\boldsymbol{y}_\xi)$ must be greater than 1, if $l_\xi = +1$ and smaller than -1, if $l_\xi = -1$. Thus, the choice of the hyperplane can be rephrased to the following optimization problem in the separable case:

$$\begin{aligned} &\text{minimize} \quad \tfrac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w} \\ &\text{subject to} \quad l_\xi(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{y} + b) \geqslant 1, \quad \xi = 1, 2, \ldots, N_{\mathscr{D}_s}. \end{aligned} \tag{29}$$

A global optimum for the parameters $\boldsymbol{w}$,b is found by using *Lagrange multipliers* (Shawe-Taylor and Cristianini, 2004). Extension to the non-separable case can be made by employing *slack variables*. The advantage of support vector classifier is that it can be extended to non-linear boundaries by the *kernel trick*. For four stress styles, the support vector classifier can achieve a correct classification rate of 46% using energy contours in several frequency bands (Fernandez and Picard, 2003).

The k-nearest neighbor classifier (k-NN) assigns an utterance to an emotional state according to the emotional state of the k utterances that are closest to u_ξ in the measurement space. In order to measure the distance between u_ξ and the neighbors, the Euclidean distance is used. The k-NN classifier achieves a correct classification rate of 64% for four emotional states using statistics of pitch and energy contours (Dellaert et al., 1996). The disadvantages of k-NN is that systematic methods for selecting the optimum number of the closest neighbors and the most suitable distance measure are hard to find. If k equals to 1, then the classifier will classify all the utterances in the design set correctly, but its performance on the test set will be poor. As $k \to \infty$, a less biased classifier is obtained. In the latter case, the optimality is not feasible for a finite number of utterances in the data collection (van der Heijden et al., 2004).

ANN-based classifiers are used for emotion classification due to their ability to find non-linear boundaries separating the emotional states. The most frequently used class of neural networks is that of feedforward ANNs, in which the input feature values propagate through the network in a forward direction on a layer-by-layer basis. Typically, the network consists of a set of sensory units that constitute the *input layer*, one or more *hidden layers* of computation nodes, and an *output layer* of computational nodes. Let us consider an one-hidden layer feedforward neural network that has Q input nodes, A hidden nodes, and B output nodes, as is depicted in Fig. 4.

The neural network provides a mapping of the form $\boldsymbol{z} = f(\boldsymbol{y})$ defined by

$$v_a = g_1(\boldsymbol{w}_a^{\mathrm{T}}\boldsymbol{y} + w_0), \tag{30}$$

$$z_{\mathrm{b}} = g_2(\boldsymbol{u}_{\mathrm{b}}^{\mathrm{T}}\boldsymbol{v} + u_0), \tag{31}$$

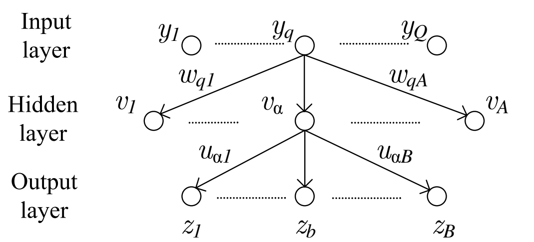

Fig. 4. An one hidden layer feedforward neural network.

where $\boldsymbol{W} = [w_{qa}] = [\boldsymbol{w}_1|\cdots|\boldsymbol{w}_a|\cdots|\boldsymbol{w}_A]$ is the weight matrix, $\boldsymbol{w}_a$ is its ath column, w_0 is the bias, and $g_1(\cdot)$ is the activation function for the input layer. Similarly $\boldsymbol{U} = [u_{ab}] = [\boldsymbol{u}_1|\ldots|\boldsymbol{u}_{\mathrm{b}}|\ldots|\boldsymbol{u}_B]$ is the weight matrix for the hidden layer, $\boldsymbol{u}_{\mathrm{b}}$ is its bth column, u_0 is the bias, and $g_2(\cdot)$ is the activation function for the hidden layer. Usually, $g_1(\cdot)$ is the sigmoid function described by

$$g_1(v) = \frac{1}{1+\exp(-v)}, \tag{32}$$

and $g_2(\cdot)$ is the softmax function defined by

$$g_2(v) = \frac{\exp(v)}{\sum_{b=1}^{B}\exp(v_{\mathrm{b}})}. \tag{33}$$

Activation functions for the hidden units are needed to introduce a non-linearity into the network. The softmax function guarantees that the outputs lie between zero and one and sum to one. Thus, the outputs of a network can be interpreted as posterior probabilities for an emotional state. The weights are updated with the back-propagation learning method (Haykin, 1998). The objective of the learning method is to adjust the free parameters of the network so that the mean square error defined by a sum of squared errors between the output of the neural network and the *target* is minimized:

$$J_{\mathrm{SE}} = \frac{1}{2}\sum_{\xi=1}^{N_{\mathscr{D}_{\mathrm{s}}}}\sum_{b=1}^{B}(f_b(\boldsymbol{y}_\xi) - l_{\xi,b})^2, \tag{34}$$

where f_b denotes the value of the bth output node. The target is usually created by assigning $l_{\xi,b} = 1$, if the label of y_ξ is Ω_{b}. Otherwise, $l_{\xi b}$ is 0. In emotion classification experiments, the ANN-based classifiers are used in two ways:

- an ANN is trained to all emotional states;
- a number of ANNs is used, where each ANN is trained to a specific emotional state.

In the first case, the number of output nodes of the ANN equals the number of emotional states, whereas in the latter case each ANN has one output node. An interesting property of ANNs is that by changing the number of hidden nodes and hidden layers we control the non-linear decision boundaries between the emotional states (Haykin, 1998; van der Heijden et al., 2004). The ANN-based classifiers may achieve a correct classification rate of 50.5% for four emotional states using energy contours in several frequency bands (Fernandez and Picard, 2003) or 75% for seven emotional states using pitch and energy contour statistics of another data collection (Schüller et al., 2004).

6. Concluding remarks

In this paper, several topics have been addressed. First, a list of data collections was provided including all available information about the databases such as the kinds of emotions, the language, etc. Nevertheless, there are still some copyright problems since the material from radio or TV is held under a limited agreement with broadcasters. Furthermore, there is a need for adopting protocols such as those in (Douglas-Cowie et al., 2003; Scherer, 2003; Schröder, 2005) that address issues related to data collection. Links with standardization activities like MPEG-4 and MPEG-7 concerning the emotion states and features should be established. It is recommended the data to be distributed by organizations (like LDC or ELRA), and not by individual research organizations or project initiatives, under a reasonable fee so that the experiments reported using the specific data collections could be repeated. This is not the case with the majority of the databases reviewed in this paper, whose terms of distribution are rather unclear.

Second, our survey has been focused on feature extraction methods that are useful in emotion recognition. The most interesting features are the pitch, the formants, the short-term energy, the MFCCs, the cross-section areas, and the Teager energy operator-based features. Features that are based on voice production models have not fully been investigated (Womack and Hansen, 1996). Non-linear aspects of speech production also contribute to the emotional speech coloring. Revisiting the fundamental models of voice production is expected to boost further the performance of emotional speech classification.

Third, techniques for speech classification into emotional states have been reviewed. The classification rates reported in the related literature are not directly comparable with each other, because they were measured on different data collections by applying different experimental protocols. Therefore, besides the availability of data collections, common experimental protocols should be defined and adopted, as for example in speech/speaker recognition, biometric person authentication, etc. Launching competitions like those regularly hosted by NIST (i.e. TREC, TRECVID, FERET, etc.)

would be worth pursuing. The techniques were separated into two categories, namely the ones that exploit timing information and those ignoring any timing information. In the former category, three techniques based on ANNs and HMMs were described. There are two differences between HMM- and ANN-based classifiers. First, HMM-based classifiers require strong assumptions about the statistical characteristics of the input, such as the parameterization of the input densities as GMMs. In many cases, correlation between the features is not included. This assumption is not required for ANN-based classifiers. An ANN learns something about the correlation between the acoustic features. Second, ANNs offer a good match with discriminative objective functions. For example, it is possible to maximize discrimination between the emotional states rather than to most faithfully approximate the distributions within each class (Morgan and Bourlard, 1995). The advantage of techniques exploiting timing information is that they can be used for speech recognition as well. A topic that has not been investigated is the evolution of emotional cues through time. Such an investigation can be achieved by a classifier that uses timing information for long speech periods. Well-known discrimination classifiers that do not exploit timing information have also been reviewed. Such classifiers include the support vector machines, the Bayes classifier with the class pdfs modeled as mixtures of Gaussians, the k-nearest neighbors, etc. The techniques that model feature pdfs may reveal cues about the modalities of the speech, such as the speaker gender and the speaker identities. One of the major drawbacks of these approaches is the *loss of the timing information*, because the techniques employ statistics of the prosody features such as the mean, the variance, etc. and neglect the sampling order. A way to overcome the problem is to calculate statistics over rising/falling slopes or during the plateaux at minima/maxima (McGilloway et al., 2000; Ververidis and Kotropoulos, 2005). It appears that most of the contour statistics follow the Gaussian distribution or the $\mathcal{X}^2$, or can be modeled by mixture of Gaussians. However, an analytical study of the feature distributions has not been undertaken yet.

Most of the emotion research activity has been focused on advancing the emotion classification performance. In spite of the extensive research in emotion recognition, efficient speech normalization techniques that exploit the emotional state information to improve speech recognition have not been developed yet.

Acknowledgment

This work has been supported by the research project 01ED312 "Use of Virtual Reality for training pupils to deal with earthquakes" financed by the Greek Secretariat of Research and Technology.

References

Abelin, A., Allwood, J., 2000. Cross linguistic interpretation of emotional prosody. In: Proc. ISCA Workshop on Speech and Emotion, Vol. 1, pp. 110–113.

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Automat. Contr. 19 (6), 716–723.

Alpert, M., Pouget, E.R., Silva, R.R., 2001. Reflections of depression in acoustic measures of the patients speech. J. Affect. Disord. 66, 59–69.

Alter, K., Rank, E., Kotz, S.A., 2000. Accentuation and emotions – two different systems? In: Proc. ISCA Workshop on Speech and Emotion, Belfast, Vol. 1, pp. 138–142.

Ambrus, D.C., 2000. Collecting and recording of an emotional speech database. Tech. rep., Faculty of Electrical Engineering, Institute of Electronics, Univ. of Maribor.

Amir, N., Ron, S., Laor, N., 2000. Analysis of an emotional speech corpus in Hebrew based on objective criteria. In: Proc. ISCA Workshop on Speech and Emotion, Belfast, Vol. 1, pp. 29–33.

Ang, J., Dhillon, R., Krupski, A., Shriberg, E., Stolcke, A., 2002. Prosody-based automatic detection of annoyance and frustration in human–computer dialog. In: Proc. Internat. Conf. on Spoken Language Processing (ICSLP '02), Vol. 3, pp. 2037–2040.

Anscombe, E., Geach, P.T. (Eds.), 1970. Descartes Philosophical Writings, second ed. Nelson, Melbourne, Australia (original work published in 1952).

Atal, B., Schroeder, M., 1967. Predictive coding of speech signals. In: Proc. Conf. on Communications and Processing, pp. 360–361.

Banse, R., Scherer, K., 1996. Acoustic profiles in vocal emotion expression. J. Pers. Soc. Psychol. 70 (3), 614–636.

Bänziger, T., Scherer, K., 2005. The role of intonation in emotional expressions. Speech Comm. 46, 252–267.

Batliner, A., Hacker, C., Steidl, S., Nöth, E., D' Archy, S., Russell, M., Wong, M., 2004. "You stupid tin box" – children interacting with the AIBO robot: a cross-linguistic emotional speech corpus. In: Proc. Language Resources and Evaluation (LREC '04), Lisbon.

Bou-Ghazale, S.E., Hansen, J., 1998. Hmm based stressed speech modelling with application to improved synthesis and recognition of isolated speech under stress. IEEE Trans. Speech Audio Processing 6, 201–216.

Buck, R., 1999. The biological affects, a typology. Psychol. Rev. 106 (2), 301–336.

Bulut, M., Narayanan, S.S., Sydral, A.K., 2002. Expressive speech synthesis using a concatenative synthesizer. In: Proc. Internat. Conf. on Spoken Language Processing (ICSLP '02), Vol. 2, pp. 1265–1268.

Burkhardt, F., Sendlmeier, W.F., 2000. Verification of acoustical correlates of emotional speech using formant-synthesis. In: Proc. ISCA Workshop on Speech and Emotion, Belfast, Vol. 1, pp. 151–156.

Cairns, D., Hansen, J.H.L., 1994. Nonlinear analysis and detection of speech under stressed conditions. J. Acoust. Soc. Am. 96 (6), 3392–3400.

Caldognetto, E.M., Cosi, P., Drioli, C., Tisato, G., Cavicchio, F., 2004. Modifications of phonetic labial targets in emotive speech: effects of the co-production of speech and emotions. Speech Comm. 44, 173–185.

Choukri, K., 2003. European Language Resources Association, (ELRA). Available from: <www.elra.info>.

Chuang, Z.J., Wu, C.H., 2002. Emotion recognition from textual input using an emotional semantic network. In: Proc. Internat. Conf. on Spoken Language Processing (ICSLP '02), Vol. 3, pp. 2033–2036.

Clavel, C., Vasilescu, I., Devillers, L., Ehrette, T., 2004. Fiction database for emotion detection in abnormal situations. In: Proc. Internat. Conf. on Spoken Language Process (ICSLP '04), Korea, pp. 2277–2280.

Cole, R., 2005. The CU kids' speech corpus. The Center for Spoken Language Research (CSLR). Available from: <http://cslr.colorado.edu/>.

Cowie, R., Cornelius, R.R., 2003. Describing the emotional states that are expressed in speech. Speech Comm. 40 (1), 5–32.

Cowie, R., Douglas-Cowie, E., 1996. Automatic statistical analysis of the signal and prosodic signs of emotion in speech. In: Proc. Internat. Conf. on Spoken Language Processing (ICSLP '96), Vol. 3, pp. 1989–1992.

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 Processing Mag. 18 (1), 32–80.

Davis, S.B., Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. Speech Signal Processing 28, 357–366.

Dellaert, F., Polzin, T., Waibel, A., 1996. Recognizing emotion in speech. In: Proc. Internat. Conf. on Spoken Language Processing (ICSLP '96), Vol. 3, pp. 1970–1973.

Deller, J.R., Hansen, J.H.L., Proakis, J.G., 2000. Discete-Time Processing of Speech Signals. Wiley, NY.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the em algorithm. J. Roy. Statist. Soc. Ser. B 39, 1–88.

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

Eckman, P., 1992. An argument for basic emotions. Cognition Emotion 6, 169–200.

Edgington, M., 1997. Investigating the limitations of concatenative synthesis. In: Proc. European Conf. on Speech Communication and Technology (Eurospeech '97), Vol. 1, pp. 593–596.

Efron, B., Tibshirani, R.E., 1993. An Introduction to the Bootstrap. Chapman & Hall/CRC, NY.

Engberg, I.S., Hansen, A.V., 1996. Documentation of the Danish Emotional Speech database (DES). Internal AAU report, Center for Person Kommunikation, Aalborg Univ., Denmark.

Fernandez, R., Picard, R., 2003. Modeling drivers' speech under stress. Speech Comm. 40, 145–159.

Fischer, K., 1999. Annotating emotional language data. Tech. Rep. 236, Univ. of Hamburg.

Flanagan, J.L., 1972. Speech Analysis, Synthesis and Perception. second ed.. Springer-Verlag, NY.

France, D.J., Shiavi, R.G., Silverman, S., Silverman, M., Wilkes, M., 2000. Acoustical properties of speech as indicators of depression and suicidal risk. IEEE Trans. Biomed. Eng. 7, 829–837.

Fukunaga, K., 1990. Introduction to Statistical Pattern Recognition. second ed.. Academic Press, NY.

Gonzalez, G.M., 1999. Bilingual computer-assisted psychological assessment: an innovative approach for screening depression in Chicanos/Latinos. Tech. Rep. 39, Univ. Michigan.

Hansen, J.H.L., 1996. NATO IST-03 (formerly RSG. 10) speech under stress web page. Available from: <http://cslr.colorado.edu/rspl/stress.html>.

Hansen, J.H.L., Cairns, D.A., 1995. ICARUS: Source generator based real-time recognition of speech in noisy stressful and Lombard effect environments. Speech Comm. 16, 391–422.

Hanson, H.M., Maragos, P., Potamianos, A., 1994. A system for finding speech formants and modulations via energy separation. IEEE Trans. Speech Audio Processing 2 (3), 436–442.

Haykin, S., 1998. Neural Networks: A Comprehensive Foundation. second ed.. Prentice Hall, NJ.

Hess, W.J., 1992. Pitch and voicing determination. In: Furui, S., Sondhi, M.M. (Eds.), Advances in Speech Signal Processing. Marcel Dekker, NY.

Heuft, B., Portele, T., Rauth, M., 1996. Emotions in time domain synthesis. In: Proc. Internat. Conf. on Spoken Language Processing (ICSLP '96), Vol. 3, pp. 1974–1977.

Iida, A., Campbell, N., Iga, S., Higuchi, F., Yasumura, M., 2000. A speech synthesis system with emotion for assisting communication. In: Proc. ISCA Workshop on Speech and Emotion, Belfast, Vol. 1, pp. 167–172.

Iida, A., Campbell, N., Higuchi, F., Yasumura, M., 2003. A corpus-based speech synthesis system with emotion. Speech Comm. 40, 161–187.

Iriondo, I., Guaus, R., Rodriguez, A., 2000. Validation of an acoustical modeling of emotional expression in Spanish using speech synthesis techniques. In: Proc. ISCA Workshop on Speech and Emotion, Belfast, Vol. 1, pp. 161–166.

Jiang, D.N., Cai, L.H., 2004. Speech emotion classification with the combination of statistic features and temporal features. In: Proc. Internat. Conf. on Multimedia and Expo (ICME '04), Taipei.

Kadambe, S., Boudreaux-Bartels, G.F., 1992. Application of the wavelet transform for pitch detection of signals. IEEE Trans. Inform. Theory 38 (2), 917–924.

Kawanami, H., Iwami, Y., Toda, T., Shikano, K., 2003. GMM-based voice conversion applied to emotional speech synthesis. In: Proc. European Conf. on Speech Communication and Technology (Eurospeech '03), Vol. 4, pp. 2401–2404.

Kwon, O.W., Chan, K.L., Hao, J., Lee, T.W., 2003. Emotion recognition by speech signals. In: Proc. European Conf. on Speech Communication and Technology (Eurospeech '03), Vol. 1, pp. 125–128.

Lee, C.M., Narayanan, S.S., 2005. Toward detecting emotions in spoken dialogs. IEEE Trans. Speech Audio Process. 13 (2), 293–303.

Leinonen, L., Hiltunen, T., Linnankoski, I., Laakso, M., 1997. Expression of emotional motivational connotations with a one-word utterance. J. Acoust. Soc. Am. 102 (3), 1853–1863.

Liberman, M., 2005. Linguistic Data Consurtium (LDC). Available from: <http://www.ldc.upenn.edu/>.

Linnankoski, I., Leinonen, L., Vihla, M., Laakso, M., Carlson, S., 2005. Conveyance of emotional connotations by a single word in English. Speech Comm. 45, 27–39.

Lloyd, A.J., 1999. Comprehension of prosody in Parkinson's disease. Proc. Cortex 35 (3), 389–402.

Makarova, V., Petrushin, V.A., 2002. RUSLANA: A database of Russian emotional utterances. In: Proc. Internat. Conf. on Spoken Language Processing (ICSLP '02), Vol. 1, pp. 2041–2044.

Mallat, S.G., Zhong, S., 1989. Complete signal representation with multiscale edges. Tech. rep., Courant Inst. of Math. Sci., rRT-483-RR-219.

Markel, J.D., Gray, A.H., 1976. Linear Prediction of Speech. Springer-Verlag, NY.

Martins, C., Mascarenhas, I., Meinedo, H., Oliveira, L., Neto, J., Ribeiro, C., Trancoso, I., Viana, C., 1998. Spoken language corpora for speech recognition and synthesis in European Portuguese. In: Proc. Tenth Portuguese Conf. on Pattern Recognition (RECPAD '98), Lisboa.

McGilloway, S., Cowie, R., Douglas-Cowie, E., Gielen, C.C.A.M., Westerdijk, M.J.D., Stroeve, S. H., 2000. Approaching automatic recognition of emotion from voice: a rough benchmark. In: Proc. ISCA Workshop on Speech and Emotion, Vol. 1, pp. 207–212.

McMahon, E., Cowie, R., Kasderidis, S., Taylor, J., Kollias, S., 2003. What chance that a DC could recognise hazardous mental states from sensor outputs? In: Tales of the Disappearing Computer, Santorini, Greece.

Mermelstein, P., 1975. Automatic segmentation of speech into syllabic units. J. Acoust. Soc. Am. 58 (4), 880–883.

Montanari, S., Yildirim, S., Andersen, E., Narayanan, S., 2004. Reference marking in children's computed-directed speech: an integrated analysis of discourse and gestures. In: Proc. Internat. Conf. on Spoken Language Processing (ICSLP '04), Korea, Vol. 1, pp. 1841–1844.

Montero, J.M., Gutierrez-Arriola, J., Colas, J., Enriquez, E., Pardo, J.M., 1999. Analysis and modelling of emotional speech in Spanish. In: Proc. Internat. Conf. on Phonetics and Speech (ICPhS '99), San Francisco, Vol. 2, pp. 957–960.

Morgan, N., Bourlard, H., 1995. Continuous speech recognition. IEEE Signal Processing Mag. 12 (3), 24–42.

Mozziconacci, S.J.L., Hermes, D.J., 1997. A study of intonation patterns in speech expressing emotion or attitude: production and perception. Tech. Rep. 32, Eindhoven, IPO Annual Progress Report.

Mozziconacci, S.J.L., Hermes, D.J., 2000. Expression of emotion and attitude through temporal speech variations. In: Proc. Internat. Conf. on Spoken Language Processing (ICSLP '00), Beijing, Vol. 2, pp. 373–378.

Mrayati, M., Carre, R., Guerin, B., 1988. Distinctive regions and models: a new theory of speech production. Speech Comm. 7 (3), 257–286.

Murray, I., Arnott, J.L., 1996. Synthesizing emotions in speech: is it time to get excited? In: Proc. Internat. Conf. on Spoken Language Processing (ICSLP '96), Vol. 3, pp. 1816–1819.

Nakatsu, R., Solomides, A., Tosa, N., 1999. Emotion recognition and its application to computer agents with spontaneous interactive capabilities. In: Proc. Internat. Conf. on Multimedia Computing and Systems (ICMCS '99), Florence, Vol. 2, pp. 804–808.

Niimi, Y., Kasamatu, M., Nishimoto, T., Araki, M., 2001. Synthesis of emotional speech using prosodically balanced VCV segments. In: Proc. ISCA Tutorial and Workshop on Research Synthesis (SSW 4), Scotland.

Nogueiras, A., Marino, J.B., Moreno, A., Bonafonte, A., 2001. Speech emotion recognition using hidden Markov models. In: Proc. European Conf. on Speech Communication and Technology (Eurospeech '01), Denmark.

Nordstrand, M., Svanfeldt, G., Granström, B., House, D., 2004. Measurements of ariculatory variation in expressive speech for a set of Swedish vowels. Speech Comm. 44, 187–196.

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

Pantic, M., Rothkrantz, L.J.M., 2003. Toward an affect-sensitive multimodal human–computer interaction. Proc. IEEE 91 (9), 1370–1390.

Pellom, B.L., Hansen, J.H.L., 1996. Text-directed speech enhancement using phoneme classification and feature map constrained vector quantization. In: Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP '96), Vol. 2, pp. 645–648.

Pereira, C., 2000. Dimensions of emotional meaning in speech. In: Proc. ISCA Workshop on Speech and Emotion, Belfast, Vol. 1, pp. 25–28.

Petrushin, V.A., 1999. Emotion in speech recognition and application to call centers. In: Proc. Artificial Neural Networks in Engineering (ANNIE '99), Vol. 1, pp. 7–10.

Picard, R.W., Vyzas, E., Healey, J., 2001. Toward machine emotional intelligence: analysis of affective physiological state. IEEE Trans. Pattern Anal. Machine Intell. 23 (10), 1175–1191.

Pollerman, B.Z., Archinard, M., 2002. Improvements in Speech Synthesis. John Wiley & Sons Ltd., England.

Polzin, T., Waibel, A., 2000. Emotion-sensitive human–computer interfaces. In: Proc. ISCA Workshop on Speech and Emotion, Belfast, Vol. 1, pp. 201–206.

Polzin, T.S., Waibel, A.H., 1998. Detecting emotions in speech. In: Proc. Cooperative Multimodal Communication (CMC '98).

Quatieri, T.F., 2002. Discrete-Time Speech Signal Processing. Prentice-Hall, NJ.

Rabiner, L.R., Juang, B.H., 1993. Fundamentals of Speech Recognition. Prentice-Hall, NJ.

Rahurkar, M., Hansen, J.H.L., 2002. Frequency band analysis for stress detection using a Teager energy operator based feature. In: Proc. Internat. Conf. on Spoken Language Processing (ICSLP '02), Vol. 3, pp. 2021–2024.

Scherer, K.R., 2000a. A cross-cultural investigation of emotion inferences from voice and speech: implications for speech technology. In: Proc. Internat. Conf. on Spoken Language Processing (ICSLP '00), Vol. 1, pp. 379–382.

Scherer, K.R., 2000b. Emotion effects on voice and speech: paradigms and approaches to evaluation. In: Proc. ISCA Workshop on Speech and Emotion, Belfast, invited paper.

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

Scherer, K.R., Banse, R., Wallbot, H.G., Goldbeck, T., 1991. Vocal clues in emotion encoding and decoding. In: Proc. Motiv. Emotion, Vol. 15, pp. 123–148.

Scherer, K.R., Grandjean, D., Johnstone, L.T., G. Klasmeyer, T.B., 2002. Acoustic correlates of task load and stress. In:

Proc. Internat. Conf. on Spoken Language Processing (ICSLP '02), Colorado, Vol. 3, pp. 2017–2020.

Schiel, F., Steininger, S., Turk, U., 2002. The Smartkom multimodal corpus at BAS. In: Proc. Language Resources and Evaluation (LREC '02).

Schröder, M., 2000. Experimental study of affect bursts. In: Proc. ISCA Workshop on Speech and Emotion, Vol. 1, pp. 132–137.

Schröder, M., 2005. Humaine consortium: research on emotions and human–machine interaction. Available from: <http://emotion-research.net/>.

Schröder, M., Grice, M., 2003. Expressing vocal effort in concatenative synthesis. In: Proc. Internat. Conf. on Phonetic Sciences (ICPhS '03), Barcelona.

Schüller, B., Rigoll, G., Lang, M., 2004. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In: Proc. Internat. Conf. on Acoustics, Speech and Signal Processing (ICASSP '04), Vol. 1, pp. 557–560.

Shawe-Taylor, J., Cristianini, N., 2004. Kernel Methods for Pattern Analysis. University Press, Cambridge.

Shi, R.P., Adelhardt, J., Zeissler, V., Batliner, A., Frank, C., Nöth, E., Niemann, H., 2003. Using speech and gesture to explore user states in multimodal dialogue systems. In: Proc. ISCA Tutorial and Research Workshop on Audio Visual Speech Processing (AVSP '03), Vol. 1, pp. 151–156.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., Hirschberg, J., 1992. ToBI: A standard for labeling English prosody. In: Proc. Internat. Conf. on Spoken Language Processing (ICSLP '92), Vol. 2, pp. 867–870.

Slaney, M., McRoberts, G., 2003. Babyears: A recognition system for affective vocalizations. Speech Comm. 39, 367–384.

Sondhi, M.M., 1968. New methods of pitch extraction. IEEE Trans. Audio Electroacoust. 16, 262–266.

Steeneken, H.J.M., Hansen, J.H.L., 1999. Speech under stress conditions: overview of the effect of speech production and on system performance. In: Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP '99), Phoenix, Vol. 4, pp. 2079–2082.

Stibbard, R., 2000. Automated extraction of ToBI annotation data from the Reading/Leeds emotional speech corpus. In: Proc. ISCA Workshop on Speech and Emotion, Belfast, Vol. 1, pp. 60–65.

Tato, R., 2002. Emotional space improves emotion recognition. In: Proc. Internat. Conf. on Spoken Language Processing (ICSLP '02), Colorado, Vol. 3, pp. 2029–2032.

Teager, H.M., Teager, S.M., 1990. Evidence for nonlinear sound production mechanisms in the vocal tractNATO Advanced Study Institute, Series D, Vol. 15. Kluwer, Boston, MA.

Tolkmitt, F.J., Scherer, K.R., 1986. Effect of experimentally induced stress on vocal parameters. J. Exp. Psychol. [Hum. Percept.] 12 (3), 302–313.

Van Bezooijen, R., 1984. The Characteristics and Recognizability of Vocal Expression of Emotions. Foris, Drodrecht, The Netherlands.

van der Heijden, F., Duin, R.P.W., de Ridder, D., Tax, D.M. J., 2004. Classification, Parameter Estimation and State Estimation – An Engineering Approach using Matlab. J. Wiley & Sons, London, UK.

Ververidis, D., Kotropoulos, C., 2004. Automatic speech classification to five emotional states based on gender information. In: Proc. European Signal Processing Conf. (EUSIPCO '04), Vol. 1, pp. 341–344.

Ververidis, D., Kotropoulos, C., 2005. Emotional speech classification using Gaussian mixture models and the sequential floating forward selection algorithm. In: Proc. Internat. Conf. on Multimedia and Expo (ICME '05).

Ververidis, D., Kotropoulos, C., Pitas, I., 2004. Automatic emotional speech classification. In: Proc. Internat. Conf. on Acoustics, Speech and Signal Processing (ICASSP '04), Montreal, Vol. 1, pp. 593–596.

Wagner, J., Kim, J., André, E., 2005. From physiological signals to emotions: implementing and comparing selected methods for feature extraction and classification. In: Proc. Internat. Conf. on Multimedia and Expo (ICME '05), Amsterdam.

Wendt, B., Scheich, H., 2002. The Magdeburger prosodiekorpus. In: Proc. Speech Prosody Conf., pp. 699–701.

Womack, B.D., Hansen, J.H.L., 1996. Classification of speech under stress using target driven features. Speech Comm. 20, 131–150.

Womack, B.D., Hansen, J.H.L., 1999. N-channel hidden Markov models for combined stressed speech classification and recognition. IEEE Trans. Speech Audio Processing 7 (6), 668–677.

Yildirim, S., Bulut, M., Lee, C.M., Kazemzadeh, A., Busso, C., Deng, Z., Lee, S., Narayanan, S., 2004. An acoustic study of emotions expressed in speech. In: Proc. Internat. Conf. on Spoken Language Processing (ICSLP '04), Korea, Vol. 1, pp. 2193–2196.

Yu, F., Chang, E., Xu, Y.Q., Shum, H.Y., 2001. Emotion detection from speech to enrich multimedia content. In: Proc. IEEE Pacific-Rim Conf. on Multimedia 2001, Beijing, Vol. 1, pp. 550–557.

Yuan, J., 2002. The acoustic realization of anger, fear, joy and sadness in Chinese. In: Proc. Internat. Conf. on Spoken Language Processing (ICSLP '02), Vol. 3, pp. 2025–2028.

Zhou, G., Hansen, J.H.L., Kaiser, J.F., 2001. Nonlinear feature based classification of speech under stress. IEEE Trans. Speech Audio Processing 9 (3), 201–216.