Pattern Recognition 44 (2011) 572–587

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Survey on speech emotion recognition: Features, classification schemes, and databases

Moataz El Ayadi [a,*], Mohamed S. Kamel [b], Fakhri Karray [b]

[a] Engineering Mathematics and Physics, Cairo University, Giza 12613, Egypt
[b] Electrical and Computer Engineering, University of Waterloo, 200 University Avenue W., Waterloo, Ontario, Canada N2L 1V9

ARTICLE INFO

Article history:
Received 4 February 2009
Received in revised form
25 July 2010
Accepted 1 September 2010

Keywords:
Archetypal emotions
Speech emotion recognition
Statistical classifiers
Dimensionality reduction techniques
Emotional speech databases

ABSTRACT

Recently, increasing attention has been directed to the study of the emotional content of speech signals, and hence, many systems have been proposed to identify the emotional content of a spoken utterance. This paper is a survey of speech emotion classification addressing three important aspects of the design of a speech emotion recognition system. The first one is the choice of suitable features for speech representation. The second issue is the design of an appropriate classification scheme and the third issue is the proper preparation of an emotional speech database for evaluating system performance. Conclusions about the performance and limitations of current speech emotion recognition systems are discussed in the last section of this survey. This section also suggests possible ways of improving speech emotion recognition systems.

1. Introduction

The speech signal is the fastest and the most natural method of communication between humans. This fact has motivated researchers to think of speech as a fast and efficient method of interaction between human and machine. However, this requires that the machine should have the sufficient intelligence to *recognize* human voices. Since the late fifties, there has been tremendous research on speech recognition, which refers to the process of converting the human speech into a sequence of words. However, despite the great progress made in speech recognition, we are still far from having a *natural* interaction between man and machine because the machine does not understand the *emotional* state of the speaker. This has introduced a relatively recent research field, namely speech emotion recognition, which is defined as extracting the emotional state of a speaker from his or her speech. It is believed that speech emotion recognition can be used to extract useful semantics from speech, and hence, improves the performance of speech recognition systems [93].

Speech emotion recognition is particularly useful for applications which require natural man–machine interaction such as web movies and computer tutorial applications where the response of those systems to the user depends on the detected emotion [116]. It is also useful for in-car board system where information of the mental state of the driver may be provided to the system to initiate his/her safety [116]. It can be also employed as a diagnostic tool for therapists [41]. It may be also useful in automatic translation systems in which the emotional state of the speaker plays an important role in communication between parties. In aircraft cockpits, it has been found that speech recognition systems trained to stressed-speech achieve better performance than those trained by normal speech [49]. Speech emotion recognition has also been used in call center applications and mobile communication [86]. The main objective of employing speech emotion recognition is to adapt the system response upon detecting frustration or annoyance in the speaker's voice.

The task of speech emotion recognition is very challenging for the following reasons. First, it is not clear which speech features are most powerful in distinguishing between emotions. The acoustic variability introduced by the existence of different sentences, speakers, speaking styles, and speaking rates adds another obstacle because these properties directly affect most of the common extracted speech features such as pitch, and energy contours [7]. Moreover, there may be more than one perceived emotion in the same utterance; each emotion corresponds to a different portion of the spoken utterance. In addition, it is very difficult to determine the boundaries between these portions. Another challenging issue is that how a certain emotion is expressed generally depends on the speaker, his or her culture and environment. Most work has focused on monolingual emotion classification, making an assumption there is no cultural difference among speakers. However, the task of multi-lingual classification has been investigated [53]. Another problem is that one may undergo a certain emotional state such as sadness for days, weeks, or even months. In such a case, other emotions will be transient and will not last for more than a few minutes. As a consequence, it is not clear which emotion

* Corresponding author.
E-mail address: moataz@pami.uwaterloo.ca (M. El Ayadi).

0031-3203/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2010.09.020

the automatic emotion recognizer will detect: the long-term emotion or the transient one. Emotion does not have a commonly agreed theoretical definition [62]. However, people know emotions when they feel them. For this reason, researchers were able to study and define different aspects of emotions. It is widely thought that emotion can be characterized in two dimensions: activation and valence [40]. Activation refers to the amount of energy required to express a certain emotion. According to some physiological studies made by Williams and Stevens [136] of the emotion production mechanism, it has been found that the sympathetic nervous system is aroused with the emotions of Joy, Anger, and Fear. This induces an increased heart rate, higher blood pressure, changes in depth of respiratory movements, greater sub-glottal pressure, dryness of the mouth, and occasional muscle tremor. The resulting speech is correspondingly loud, fast and enunciated with strong high-frequency energy, a higher average pitch, and wider pitch range. On the other hand, with the arousal of the parasympathetic nervous system, as with sadness, heart rate and blood pressure decrease and salivation increases, producing speech that is slow, low-pitched, and with little high-frequency energy. Thus, acoustic features such as the pitch, timing, voice quality, and articulation of the speech signal highly correlate with the underlying emotion [20]. However, emotions cannot be distinguished using only activation. For example, both the anger and the happiness emotions correspond to high activation but they convey different affect. This difference is characterized by the valence dimension. Unfortunately, there is no agreement within researchers on how, or even if, acoustic features correlate with this dimension [79]. Therefore, while classification between high-activation (also called high-arousal) emotions and low-activation emotions can be achieved at high accuracies, classification between different emotions is still challenging.

An important issue in speech emotion recognition is the need to determine a set of the important emotions to be classified by an automatic emotion recognizer. Linguists have defined inventories of the emotional states, most encountered in our lives. A typical set is given by Schubiger [111] and O'Connor and Arnold [95], which contains 300 emotional states. However, classifying such a large number of emotions is very difficult. Many researchers agree with the 'palette theory', which states that any emotion can be decomposed into *primary* emotions similar to the way that any color is a combination of some basic colors. Primary emotions are Anger, Disgust, Fear, Joy, Sadness, and Surprise [29]. These emotions are the most obvious and distinct emotions in our life. They are called the *archetypal* emotions [29].

In this paper, we present a comprehensive review of speech emotion recognition systems targeting pattern recognition researchers who do not necessarily have a deep background in speech analysis. We survey three important aspects in speech emotion recognition: (1) important design criteria of emotional speech corpora, (2) the impact of speech features on the classification performance of speech emotion recognition, and (3) classification systems employed in speech emotion recognition. Though there are many reviews on speech emotion recognition such as [129,5,12], our survey is more comprehensive in surveying the speech features and the classification techniques used in speech emotion recognition. We surveyed different types of features and considered the benefits of combining the available acoustic information with other sources of information such as linguistic, discourse, and video information. We theoretically covered, in some detail different classification techniques commonly used in speech emotion recognition. We also included numerous speech recognition systems implemented in other research papers in order to have an insight on the performance of existing speech emotion recognizers. However, the reader should interpret the recognition rates of those systems carefully since different emotional speech corpora and experimental setups were used with each of them.

The paper is divided into five sections. In Section 2, important issues in the design of an emotional speech database are discussed. Section 3 reviews in detail speech feature extraction methods. Classification techniques applied in speech emotion recognition are addressed in Section 4. Finally, important conclusions are drawn in Section 5.

2. Emotional speech databases

An important issue to be considered in the evaluation of an emotional speech recognizer is the degree of naturalness of the database used to assess its performance. Incorrect conclusions may be established if a low-quality database is used. Moreover, the design of the database is critically important to the classification task being considered. For example, the emotions being classified may be infant-directed; e.g. soothing and prohibition [15,120], or adult-directed; e.g. joy and anger [22,38]. In other databases, the classification task is to detect stress in speech [140]. The classification task is also defined by the number and type of emotions included in the database. This section is divided into three subsections. In Section 2.1, different criteria used to evaluate the goodness of an emotional speech database are discussed. In Section 2.2, a brief overview of some of the available databases is given. Finally, limitations of the emotional speech databases are addressed in Section 2.3.

2.1. Design criteria

There should be some criteria that can be used to judge how well a certain emotional database simulates a real-world environment. According to some studies [69,22], the following are the most relevant factors to be considered:

Real-world emotions or acted ones?: It is more realistic to use speech data that are collected from real life situations. A famous example is the recordings of the radio news broadcast of major events such as the crash of Hindenburg [22]. Such recordings contain utterances with very natural conveyed emotions. Unfortunately, there may be some legal and moral issues that prohibit the use of them for research purposes. Alternatively, emotional sentences can be elicited in sound laboratories as in the majority of the existing databases. It has always been criticized that acted emotions are not the same as real ones. Williams and Stevens [135] found that acted emotions tend to be more exaggerated than real ones. Nonetheless, the relationship between the acoustic correlate and the acted emotions does not contradict that between acoustic correlates and real ones.

Who utters the emotions?: In most emotional speech databases, professional actors are invited to express (or feign) pre-determined sentences with the required emotions. However, in some of them such as the Danish Emotional Speech (DES) database [38], semi-professional actors are employed instead in order to avoid exaggeration in expressing emotions and to be closer to real-world situations.

How to simulate the utterances?: The recorded utterances in most emotional speech databases are not produced in a conversational context [69]. Therefore, utterances may lack some naturalness since it is believed that most emotions are outcomes of our response to different situations. Generally, there are two approaches for eliciting emotional utterances. In the first approach, experienced speakers act as if they were in a specific emotional state, e.g. being glad, angry, or sad. In many developed corpora [15,38], such experienced actors were not available and semi-professional or amateur actors were invited to utter the emotional utterances. Alternatively, a Wizard-of-Oz scenario is

used in order to help the actor reach the required emotional states. This wizard involves the interaction between the actor and the computer as if the latter is a human [8]. In a recent study [59], it was proposed to use computer games to induce natural emotional speech. Voice samples were elicited following game events whether the player won or lost the game and were accompanied by either pleasant or unpleasant sounds.

Balanced utterances or unbalanced utterances?: While balanced utterances are useful for controlled scientific analysis and experiments, they may reduce the validity of the data. As an alternative, a large set of unbalanced and valid utterances may be used.

Utterances are uniformly distributed over emotions?: Some corpus developers prefer that the number of utterances for each emotion is almost the same in order to properly evaluate the classification accuracy such as in the Berlin corpus [18]. On the other hand, many other researchers prefer that the distribution of the emotions in the database reflects their frequency in the world [140,91]. For example, the neutral emotion is the most frequent emotion in our daily life. Hence, the number of utterances with neutral emotion should be the largest in the emotional speech corpus.

Same statement with different emotions?: In order to study the explicit effect of emotions on the acoustic features of the speech utterances, it is common in many databases to record the same sentence with different emotions. One advantage of such a database is to ensure that the human judgment on the perceived emotion is solely based on the emotional content of the sentence and not on its lexical content.

2.2. Available and known emotional speech databases

Most of the developed emotional speech databases are not available for public use. Thus, there are very few benchmark databases that can be shared among researchers. Another consequence from this privacy is the lack of coordination among researchers in this field: the same mistakes in recording are being repeated for different emotional speech databases. Table 1 summarizes characteristics of some databases commonly used in speech emotion recognition. From this table, we notice that the emotions are usually stimulated by professional or nonprofessional actors. In fact, there are some legal and ethical issues that may prevent researchers from recording real voices. In addition, nonprofessional actors are invited to produce emotions in many databases in order to avoid exaggeration in the perceived emotions. Moreover, we notice that most the databases share the following emotions: anger, joy, sadness, surprise, boredom, disgust, and neutral following the palette theory. Finally, most of the databases addressed adult-directed emotions while only two, KISMET and BabyEars, considered infant-directed emotions. It is believed that recognizing infant-directed emotions is very useful in the interaction between man and robots [15].

2.3. Problems in existing emotional speech databases

Almost all the existing emotional speech databases have some limitations for assessing the performance of proposed emotion recognizers. Some of the limitations of emotional speech databases are briefly mentioned:

(1) Most speech emotional databases do not well enough simulate emotions in a natural and clear way. This is evidenced by the relatively low recognition rates of human subjects. In some databases (see [94]), the human recognition performance is as low as about 65%.
(2) In some databases such as KISMET, the quality of the recorded utterances is not so good. Moreover, the sampling frequency is somewhat low (8 kHz).
(3) Phonetic transcriptions are not provided with some databases such as BabyEars [120]. Thus, it is difficult to extract linguistic content from the utterances of such databases.

3. Features for speech emotion recognition

An important issue in the design of a speech emotion recognition system is the extraction of suitable features that efficiently characterize different emotions. Since pattern recognition techniques are rarely independent of the problem domain, it is believed that a proper selection of features significantly affects the classification performance.

Four issues must be considered in feature extraction. The first issue is the region of analysis used for feature extraction. While some researchers follow the ordinary framework of dividing the speech signal into small intervals, called *frames*, from each which a local feature vector is extracted, other researchers prefer to extract global statics from the whole speech utterance. Another important question is what the best feature types for this task are, e.g. pitch, energy, zero crossing, etc.? A third question is what is the effect of ordinary speech processing such as post-filtering and silence removal on the overall performance of the classifier? Finally, whether it suffices to use acoustic features for modeling emotions or if it is necessary to combine them with other types of features such as linguistic, discourse information, or facial features.

The above issues are discussed in detail in the following five subsections. In Section 3.1, a comparison between local features and global features is given. Section 3.2 describes different types of speech features used in speech emotion recognition. This subsection is concluded with our recommendations for the choice of speech features. Section 3.3 explains the pre-processing and the post-processing steps required for the extracted speech features. Finally, Section 3.4 discusses other sources of information that can be integrated with the acoustic one in order to improve classification performance.

3.1. Local features versus global features

Since speech signals are not stationary even in wide sense, it is common in speech processing to divide a speech signal into small segments called frames. Within each frame the signal is considered to be approximately stationary [104]. Prosodic speech features such as pitch and energy are extracted from each frame and called local features. On the other hand, global features are calculated as statistics of all speech features extracted from an utterance. There has been a disagreement on which of local and global features are more suitable for speech emotion recognition. The majority of researchers have agreed that global features are superior to local ones in terms of classification accuracy and classification time [128,57,117,100]. Global features have another advantage over local features; their number is much less. Therefore, the application of cross validation and feature selection algorithms to global features are executed much faster than if applied to local features.

However, researchers have claimed that global features are efficient only in distinguishing between high-arousal emotions, e.g. anger, fear, and joy, versus low-arousal ones, e.g. sadness [94]. They claim that global features fail to classify emotions which have similar arousal, e.g. Anger versus Joy. Another disadvantage of global features is that temporal information present in speech signals is completely lost. Moreover, it may be unreliable to use complex classifiers such as the hidden Markov model (HMM) and the support vector machine (SVM) with global speech features since the number of training vectors may not be sufficient for reliably estimating model parameters. On the other hand, complex classifiers can be trained reliably using the large number of local

Table 1
Characteristics of common emotional speech databases.

Corpus	Access	Language	Size	Source	Emotions
LDC Emotional Prosody Speech and Transcripts [78]	Commercially available[a]	English	7 actors × 15 emotions × 10 utterances	Professional actors	Neutral, panic, anxiety, hot anger, cold anger, despair, sadness, elation, joy, interest, boredom, shame, pride, contempt
Berlin emotional database [18]	Public and free[b]	German	800 utterances (10 actors × 7 emotions × 10 utterances + some second version) = 800 utterances	Professional actors	Anger, joy, sadness, fear, disgust, boredom, neutral
Danish emotional database [38]	Public with license fee[c]	Danish	4 actors × 5 emotions (2 words + 9 sentences + 2 passages)	Nonprofessional actors	Anger, joy, sadness, surprise, neutral
Natural [91]	Private	Mandarin	388 utterances, 11 speakers, 2 emotions	Call centers	Anger, neutral
ESMBS [94]	Private	Mandarin	720 utterances, 12 speakers, 6 emotions	Nonprofessional actors	Anger, joy, sadness, disgust, fear, surprise
INTERFACE [54]	Commercially available[d]	English, Slovenian, Spanish, French	English (186 utterances), Slovenian (190 utterances), Spanish (184 utterances), French (175 utterances)	Actors	Anger, disgust, fear, joy, surprise, sadness, slow neutral, fast neutral
KISMET [15]	Private	American English	1002 utterances, 3 female speakers, 5 emotions	Nonprofessional actors	Approval, attention, prohibition, soothing, neutral
BabyEars [120]	Private	English	509 utterances, 12 actors (6 males + 6 females), 3 emotions	Mothers and fathers	Approval, attention, prohibition
SUSAS [140]	Public with license fee[e]	English	16,000 utterances, 32 actors (13 females + 19 males)	Speech under simulated and actual stress	Four stress styles: Simulated Stress, Calibrated Workload Tracking Task, Acquisition and Compensatory Tracking Task, Amusement Park Roller-Coaster, Helicopter Cockpit Recordings
MPEG-4 [114]	Private	English	2440 utterances, 35 speakers	U.S. American movies	Joy, anger, disgust, fear, sadness, surprise, neutral
Beihang University [43]	Private	Mandarin	7 actors × 5 emotions × 20 utterances	Nonprofessional actors	Anger, joy, sadness, disgust, surprise
FERMUS III [112]	Public with license fee[f]	German, English	2829 utterances, 7 emotions, 13 actors	Automotive environment	Anger, disgust, joy, neutral, sadness, surprise
KES [65]	Private	Korean	5400 utterances, 10 actors	Nonprofessional actors	Neutral, joy, sadness, anger
CLDC [146]	Private	Chinese	1200 utterances, 4 actors	Nonprofessional actors	Joy, anger, surprise, fear, neutral, sadness
Hao Hu et al. [56]	Private	Chinese	8 actors × 5 emotions × 40 utterances	Nonprofessional actors	Anger, fear, joy, sadness, neutral
Amir et al. [2]	Private	Hebrew	60 Hebrew and 1 Russian actors	Nonprofessional actors	Anger, disgust, fear, joy, neutral, sadness
Pereira [55]	Private	English	2 actors × 5 emotions × 8 utterances	Nonprofessional actors	Hot anger, cold anger, joy, neutral, sadness

[a] Linguistic Data Consortium, University of Pennsylvania, USA.
[b] Institute for Speech and Communication, Department of Communication Science, the Technical University, Germany.
[c] Department of Electronic Systems, Aalborg University, Denmark.
[d] Center for Language and Speech Technologies and Applications (TALP), the Technical University of Catalonia, Spain.
[e] Linguistic Data Consortium, University of Pennsylvania, USA.
[f] FERMUS research group, Institute for Human-Machine Communication, Technische Universität München, Germany.

feature vectors and hence their parameters will be accurately estimated. This may lead to higher classification accuracy than that achieved if global features are used.

A third approach for feature extraction is based on segmenting speech signals to the underlying phonemes and then calculating one feature vector for each segmented phoneme [73]. This approach relies on a study that observes variation in the spectral shapes of the same phone under different emotions [74]. This observation is essentially true for vowel sounds. However, the poor performance of phoneme segmentation algorithms can be another problem, especially when the phonetic transcriptions of utterances are not provided. An alternative method is to extract a feature vector for each voiced speech segment rather than for each phoneme. Voiced speech segments refer to continuous parts of speech that are caused by vibrations of the vocal cord and are oscillatory [104]. This approach is much easier to implement than the phoneme-based approach. In [117], the feature vector contained a combination of segment-based and global features. The *k*-nearest neighbor (*k*-NN) and the SVM were used for classification. The KISMET emotional corpus [15] was used for assessing the classification performance. The corpus contained 1002 utterances from three English speakers with the following infant-directed emotions: approval, attention, prohibition, soothing, and neutral. Speaker-dependent classification was mainly considered. Employing their feature representation resulted in 5% increase over the baseline accuracy corresponding to using only global features. In particular, the segment-based approach achieved classification accuracies of 87% and 83% using

the *k*-NN and the SVM, respectively, versus 81% and 78% obtained by utterance-level features and using the same classifiers.

3.2. Categories of speech features

An important issue in speech emotion recognition is the extraction of speech features that efficiently characterize the emotional content of speech and at the same time do not depend on the speaker or the lexical content. Although many speech features have been explored in speech emotion recognition, researchers have not identified the best speech features for this task.

Speech features can be grouped into four categories: continuous features, qualitative features, spectral features, and TEO (Teager energy operator)-based features. Fig. 1 shows examples of features belonging to each category. The main purpose of this section is to compare the pros and cons of each category. However, it is common in speech emotion recognition to combine features that belong to different categories to represent the speech signal.

3.2.1. Continuous speech features

Most researchers believe that prosody continuous features such as pitch and energy convey much of the emotional content of an utterance [29,19,12]. According to the studies performed by Williams and Stevens [136], the arousal state of the speaker (high activation versus low activation) affects the overall energy, energy distribution across the frequency spectrum and the frequency and duration of pauses of speech signal. Recently, several studies have confirmed this conclusion [60,27].

Continuous speech features have been heavily used in speech emotion recognition. For example, Banse et al. examined vocal cues for 14 emotion categories [7]. The speech features they used are related to the fundamental frequency (F0), the energy, the articulation rate, and the spectral information in voiced and unvoiced portions. According to many studies (see [29,92,69]), these acoustic features can be grouped into the following categories:

(1) pitch-related features;
(2) formants features;
(3) energy-related features;
(4) timing features;
(5) articulation features.

Some of the most commonly used global features in speech emotion recognition are:

Fundamental frequency (*F* 0): mean, median, standard deviation, maximum, minimum, range (max–min), linear regression coefficients, 4th order Legendre parameters, vibrations, mean of first difference, mean of the absolute of the first difference, jitter, and ratio of the sample number of the up-slope to that of the down-slope of the pitch contour.

Energy: mean, median, standard deviation, maximum, minimum, range (max–min), linear regression coefficients, shimmer, and 4th order Legendre parameters.

Duration: speech rate, ratio of duration of voiced and unvoiced regions, and duration of the longest voiced speech.

Formants: first and second formants, and their bandwidths.

More complex statistics are also used such as the parameters of the F0-pattern generation model proposed by Fujisaki (for more details, see [51]).

Several studies on the relationship between the above-mentioned speech features and the basic archetypal emotions have been made [28,29,7,92,96,9,11,123]. From these studies, it has been shown that prosodic features provide a reliable indication of the emotion. However, there are contradictory reports on the effect of emotions on prosodic features. For example, while Murray and Arnott [92] indicate that a high speaking rate is associated with the emotion of anger, Oster and Risberg [96] have an opposite conclusion. In addition, it seems that there are similarities between characteristics of some emotions. For instance, the emotions of anger, fear, joy, and surprise have similar characteristics for the fundamental frequency (F0) [104,20] such as:

- Average pitch: average value of F0 for the utterance.
- Contour slope: the slope of the F0-contour.
- Final lowering: the steepness of the F0 decrease at the end of the falling contour, or of the rise at the end of rising contour.
- Pitch range: the difference between the highest and the smallest value of F0.
- Reference line: the steady value of F0 after an excursion of high or small pitch.

3.2.2. Voice quality features

It is believed that the emotional content of an utterance is strongly related to its voice quality [29,109,31]. Experimental studies with listening human subjects demonstrated a strong relation between voice quality and the perceived emotion [46]. Many researchers studying the auditory aspects of emotions have been trying to define a relation [29,92,28,110]. Voice quality seems to be described most regularly with reference to full-blown emotions; i.e. emotions that strongly direct people into a course of actions [29]. This is opposed to "underlying emotions" which influence positively or negatively a person's actions and thoughts

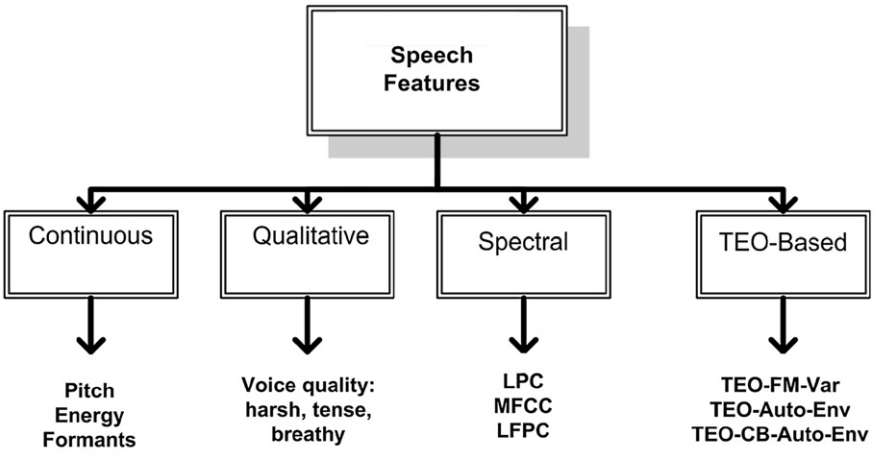

Fig. 1. Categories of speech features.

without seizing control [29]. A wide range of phonetic variables contributes to the subjective impression of voice quality [92]. According to an extensive study made by Cowie et al. [29], the acoustic correlates, related to the voice quality, are grouped into the following categories.

(1) voice level: signal amplitude, energy and duration have been shown to be reliable measures of voice level;
(2) voice pitch;
(3) phrase, phoneme, word and feature boundaries;
(4) temporal structures.

However, relatively little is known about the role of voice quality in delivering emotions for two reasons. First, impressionistic labels are used to described voice quality such as tense, harsh, and breathy. Those terms can have different interpretations based on the understanding of the researcher [46]. This led to a disagreement between researchers on how to associate vocal quality terms to emotion. For example, Sherer [109] suggested that tense voice is associated with anger, joy, and fear; and lax voice is associated with sadness. On the other hand, Murray and Arnott [92] suggested that breathy voice is associated with both anger and happiness; sadness is associated with a 'resonant' voice quality.

The second problem is the difficulty of automatically deciding those voice quality terms directly from the speech signal. There has been numerous research for the latter problem which can be categorized into two approaches. The first approach depends on the fact that the speech signal can be modelled as the output of vocal tract filter excited by a glottal source signal [104]. Therefore, voice quality can better measured by removing the filtering effect of the vocal tract and measuring parameters of the glottal signal. However, neither the glottal source signal nor the vocal tract filter are known and hence the glottal signal is estimated by exploiting knowledge about the characteristics of the source signal and of the vocal tract filter. For a review of inverse-filtering techniques, the reader is referred to [46] and the references therein. Because of the inherent difficulty in this approach, it is not much used in speech emotion recognition; e.g. [122]. In the second approach, the voice quality is numerically represented by parameters estimated directly from the speech signal; i.e. no estimation of the glottal source signal is performed. In [76], voice quality was represented by the jitter and shimmer [44]. The speech emotion recognition system used continuous HMM as a classifier and applied to utterances from the SUSAS database [140] with the following selected speaking styles: angry, fast, Lombard, question, slow and soft. The classification task was speaker independent but dialect-dependent. The baseline accuracy corresponding to using only MFCC as features was 65.5%. The classification accuracy was 68.1% when the MFCC was combined with the jitter, 68.5% when the MFCC was combined with the shimmer, and 69.1% when the MFCC was combined with both of them.

In [81,83,84], voice quality parameters are roughly calculated as follows. The pitch, the first four formant frequencies and their bandwidths are estimated from the speech signal. The effect of vocal tract is equalized mathematically by subtracting terms which represent the vocal tract influence from the amplitudes of each harmonic (see [85] for details). Finally, voice quality parameters, called spectral gradients, are calculated as simple functions of the compensated harmonic amplitudes. The experimental result of their study is discussed in Section 4.5.

3.2.3. Spectral-based speech features

In addition to time-dependent acoustic features such as pitch and energy, spectral features are often selected as a short-time representation for speech signal. It is recognized that the emotional content of an utterance has an impact on the distribution of the spectral energy across the speech range of frequency [94]. For example, it is reported that utterances with happiness emotion have high energy at high frequency range while utterances with the sadness emotion have small energy at the same range [7,64].

Spectral features can be extracted in a number of ways including the ordinary linear predictor coefficients (LPC) [104], one-sided autocorrelation linear predictor coefficients (OSALPC) [50], short-time coherence method (SMC) [14], and least-squares modified Yule–Walker equations (LSMYWE) [13]. However, in order to better exploit the spectral distribution over the audible frequency range, the estimated spectrum is often passed through a bank of band-pass filters. Spectral features are then extracted from the outputs of these filters. Since human perception of pitch does not follow a linear scale [103], the filters' bandwidths are usually evenly distributed with respect to a suitable nonlinear frequency scale such as the Bark scale [103], the Mel-frequency scale [103,61], the modified Mel-frequency scale, and the ExpoLog scale [13].

Cepstral-based features can be derived from the corresponding linear features as in the case of linear predictor cepstral coefficients (LPCC) [4] and cepstral-based OSALPC (OSALPCC) [13]. There have been contradictory reports on whether cepstral-based features are better than linear-based ones in emotion recognition. In [13], it was shown that features based on cepstral analysis such as LPCC, OSALPCC, and Mel-frequency cepstrum coefficients (MFCC) clearly outperform the performance of the linear-based features of LPC and OSALPC, in detecting stress in speech signal. However, New et al. [94] compared a linear-based feature, namely Log-frequency power coefficients (LFPC), and two cepstral-based features, namely LPCC and MFCC. They mainly used HMM for classification. The emotional speech database they used was locally recorded. It contained 720 utterances from six Burmese speakers and six Mandarin speakers with the six archetypal emotions: anger, disgust, fear, joy, sadness, and surprise. Sixty percent of the emotion utterances of each speaker were used to train each emotion model while the remaining 40% of the utterances were used for testing. They showed that the LFPC provided an average classification accuracy of 77.1% while the LPCC and the MFCC gave 56.1% and 59.0% identification accuracies, respectively.

3.2.4. Nonlinear TEO-based features

According to experimental studies done by Teager, the speech is produced by nonlinear air flow in the vocal system [125]. Under stressful conditions, the muscle tension of the speaker affects the air flow in the vocal system producing the sound. Therefore, nonlinear speech features are necessary for detecting the speech in the sound. The Teager-energy-operator (TEO), first introduced by Teager [124] and Kaiser [63], was originally developed with the supporting evidence that hearing is the process of detecting energy. For a discrete time signal, $x[n]$, the TEO is defined as

$$\Psi\{x[n]\} = x^2[n] - x[n-1]x[n+1]. \quad (1)$$

It has been observed that under stressful conditions the fundamental frequency changes, as does the distribution of harmonics over the critical bands [13,125]. It is verified that the TEO of multi-frequency signal does not only reflects individual frequency components but also interaction between them [145]. Based on this fact, TEO-based features can then be used for detecting stress in speech. In [21], the Teager energy profile of the pitch contour was the feature used to classify the following effects in speech: loud, angry, Lombard, clear, and neutral. Classification was performed by a combination of vector quantization and HMM. The classification system was applied to utterances from the SUSAS database [140] and it was speaker

dependent. While the classification system detected the loud and angry effects of speech with a high accuracy of 98.1% and 99.1%, respectively, the classification accuracies of detecting the Lombard and clear effects were much lower: 86.1% and 64.8%. Moreover, two assumptions were made: (1) the text of the spoken utterances is already known to the system, and (2) the spoken words have the structure of vowel-consonant or consonant-vowel-consonant. Therefore, much lower accuracies are expected for free-style speech.

In another study [145], other TEO-based features, namely TEO-decomposed FM variation (TEO-FM-Var), normalized TEO autocorrelation envelope area (TEO-Auto-Env), and critical band-based TEO autocorrelation envelope area (TEO-CB-Auto-Env), were proposed for detecting neutral versus stressed speech and for classifying the stressed speech into three styles: angry, loud, and Lombard. Five-state HMM was used as a baseline classifier and tested using utterances from the SUSAS database [140]. The developed features were compared against the MFCC and the pitch features in three classification tasks:

(1) Text-dependent pairwise stress classification:
TEO-FM-Var (70.5% $\pm$ 15.77%), TEO-Auto-Env (79.4% $\pm$ 4.01%), TEO-CB-Auto-Env (92.9% $\pm$ 3.97%), MFCC (90.9% $\pm$ 5.73%), Pitch (79.9% $\pm$ 17.18%).
(2) Text-independent pairwise stress classification:
TEO-CB-Auto-Env (89.0% $\pm$ 8.39%), MFCC (67.7% $\pm$ 8.78%), Pitch (79.9% $\pm$ 17.18%).
(3) Text-independent multi-style stress classification:
TEO-CB-Auto-Env (Neutral 70.6%, Angry 65.0%, Loud 51.9%, Lombard 44.9%), MFCC (Neutral 46.3%, Angry 58.6%, Loud 20.7%, Lombard 35.1%), Pitch (Neutral 52.2%, Angry 44.4%, Loud 53.3%, Lombard 89.5%).

Based on the extensive experimental evaluations, the authors concluded that TEO-CB-Auto-Env outperformed the MFCC and the pitch in stress detection but it completely fails for the composite task of speech recognition and stress classification.

We also conclude that the choice of proper features for speech emotion recognition highly depends on the classification task being considered. In particular, based on the review in this section, we recommend the use of TEO-based features for detecting stress in speech. For classifying high-arousal versus low-arousal emotions, continuous features such as the fundamental frequency and the pitch should be used. For *N*-way classification, the spectral features such as the MFCC are the most promising features for speech representation. We also believe that combining continuous and spectral features will provide even a better classification performance for the same task. Clearly there are some relationships among the feature types described above. For example, spectral variables relate to voice quality, and the pitch contours relate to the patterns arising from different tones. But links are rarely made in the literature.

3.3. Speech processing

The term *pre-processing* refers to all operations, required to be performed on the time samples of speech signal before extracting features. For example, due to recording environment differences, some sort of energy normalization has to be done to all utterances. In order to equalize the effect of the propagation of speech through air, a pre-emphasis radiation filter is used to process speech signal before extraction of features. The transfer function of the pre-emphasis filter is usually given by [104]

$$H(z) = 1 - 0.97z^{-1}. \quad (2)$$

In order to smooth the extracted contours, overlapped frames are commonly used. In addition, to reduce ripples in the spectrum of the speech spectrum, each frame is often multiplied by a Hamming window before feature extraction [104].

Since the silence intervals carry important information about the expressed emotion [94], these intervals are usually kept intact in speech emotion recognition. Note that silent intervals are frequently omitted from analysis in other spoken language tasks, such as speaker identification [107].

Having extracted the suitable speech features from the pre-processed time samples, some *post-processing* may be necessary before the feature vectors are used to train or test the classifier. For example, the extracted features may be of different units and hence their numerical values have different orders of magnitude. In addition, some of them may be biased. This can cause some numerical problems in training some classifiers, e.g. the Gaussian mixture model (GMM), since the covariance matrix of the training data may be ill conditioned. Therefore, feature normalization may be necessary in such cases. The most common method for feature normalization is through *z*-score normalization [116,115]:

$$\hat{x} = \frac{x - \mu}{\sigma}, \quad (3)$$

where μ is the mean of the feature x and σ is the standard deviation. However, a disadvantage of this method is that all the normalized features have a unity variance. It is believed that the variances of features have high information content [90].

It is also common to use dimensionality reduction techniques in speech emotion recognition applications in order to reduce the storage and computation requirements of the classifier and to have an insight about the discriminating features. There are two approaches for dimensionality reduction: feature selection and feature extraction (also called feature transform [80]). In feature selection, the main objective is to find the feature subset that achieves the best possible classification between classes. The classification ability of a feature subset is usually characterized by an easy-to-calculate function, called the feature selection criterion, such as the cross validation error [10] and the mutual information between the class label and the feature [137]. On the other hand, feature extraction techniques aims at finding a suitable linear or nonlinear mapping from the original feature space to another space with reduced dimensionality while preserving as much relevant classification information as possible. The reader may refer to [58,34] for excellent reviews on dimensionality reduction techniques.

The principle component analysis (PCA) feature extraction method has been used extensively in the context of speech emotion recognition [141,143,130,71]. In [25], it is observed that increasing the number of principle components improves the classification performance until a certain order after which the classification accuracy begins to decrease. This means that employing PCA may provide an improvement in the classification performance over using the whole feature set. It is not clear also whether the PCA is superior to other dimensionality reduction techniques. While the performance of the PCA was very comparable to the linear discriminant analysis (LDA) in [143], it is reported in [141,142] that the PCA is significantly inferior to the LDA and the sequential floating search (SFS) dimensionality techniques. The obvious interpretation is that different acoustic features and emotional databases are used in those studies.

The LDA has also been applied in speech emotion recognition applications [141,143] though it has the limitation that the reduced dimensionality must be less than the number of classes [34]. In [116], the LDA technique is used to compare more than 200 speech features. According to this study, it is concluded that pitch-related features yield about 69.81% recognition accuracy versus 36.58%, provided by energy-related features. This result is opposed to that established in [101] where it is concluded that the first and third

quartiles in the energy distribution are important features in the task of emotion classification. In order to establish a reliable conclusion about a certain feature as being powerful in distinguishing different emotional classes, one has to do ranking over more than one database.

3.4. Combining acoustic features with other information sources

In many situations, nonacoustic emotional cues such as facial expressions or some specific words are helpful to understand the desired speaker's emotion. This fact has motivated some researchers to employ other sources of information in conjunction with the acoustic correlates in order to improve the recognition performance. In this section, a detailed overview of some emotion recognition systems that apply this idea is presented.

3.4.1. Combining acoustic and linguistic information

Linguistic content of the spoken utterance is an important part of the conveyed emotion [33]. Recently, there has been a focus on the integration of acoustic and linguistic information [72]. In order to make use of the linguistic information, it is first necessary to recognize the word sequence of the spoken utterance. Therefore, a language model is necessary. Language models describe constraints on possible word sequences in a certain language. A common language model is the N-gram model [144]. This model assigns high probabilities to typical word sequences and low probabilities for atypical word sequences [5].

Fig. 2 shows the basic architecture of a speech emotion recognition that combines the roles of acoustic and linguistic models in finding the most probable word sequence. The input word-transcriptions are processed in order to produce the language model.[1] In parallel, the feature extraction module converts speech signal into a sequence of feature vectors. The extracted feature vectors together with the pronunciation dictionary and the input word-transcriptions are then used to train the phoneme acoustic models. In the recognition phase, both the language model and the acoustic models obtained in the training phase are used to recognize the output word sequence according to the following Bayes rule:

$$\hat{W} = \arg\max_{W} P(W|Y) = \arg\max_{W} \frac{P(W)P(Y|W)}{P(Y)} = \arg\max_{W} P(W)P(Y|W), \quad (4)$$

where Y is the set of acoustic feature vectors produced by the feature extraction module. The prior word probability is determined directly from the language model. In order to estimate the conditional probability of the acoustic feature set given a certain word sequence, a HMM for each phoneme is constructed and trained based on available speech database. The required conditional probability is estimated as the likelihood value produced by a set of phoneme HMMs concatenated in a sequence according to the word transcription stored in the dictionary. The Viterbi algorithm [131] is usually used for searching for the optimum word sequence that produced the given testing utterances.

In [116], a spotting algorithm that searches for emotional keywords or phrases in the utterances was employed. A Bayesian belief network was used to recognize emotions based on the acoustic features extracted from these keywords. The emotional speech corpus was collected from the FERMUS III project [112]. The corpus contained the following emotions: angry, disgust, fear, joy, neutral, sad, and surprise. The k-means, the GMM, the multi-layer perceptron (MLP), and the SVM classifiers were used to classify emotions based on the acoustic information. The SVM provided the best speaker-independent classification accuracy (81.29%) and thus selected as the acoustic classifier to be integrated with the linguistic classifier. The decisions of the acoustic and linguistic classifiers were fused by a MLP neural network. In that study, it was shown that the average recognition accuracy was 74.2% for acoustic features alone, 59.6% for linguistic information alone, 83.1% for both acoustic and linguistic using fusion by mean and 92.0% for fusion by MLP neural network.

An alternative procedure for detecting emotions using lexical information is found in [69]. In this work, a new information theoretic measure, named emotional salience, was defined. Emotional salience measures how much information a word provides towards a certain emotion. This measure is more or less related to the mutual information between a particular word and a certain emotional category [47]. The training data set was selected 10 times in a random manner from the whole data set for each gender with the same number of data for each class (200 for male data and 240 for female). Using acoustic information only, the classification error ranged from 17.85% to 25.45% for male data and from 12.04% to 24.25% for female data. The increase in the classification accuracy due to combining the linguistic information with the acoustic information was in the range from 7.3% to 11.05% for male data and 4.05% to 9.47% for female data.

3.4.2. Combining acoustic, linguistic, and discourse information

Discourse markers are linguistic expressions that convey explicit information about the structure of the discourse or have a specific semantic contribution [48,26]. In the context of speech emotion recognition, discourse information may also refer to the way a user interacts with the machine [69]. Often, these systems do not operate in a perfect manner; and hence, it might happen that the user expresses some emotion such as frustration in response to them [3]. Therefore, it is believed that there is a strong relation between the way a user interacts with a system and his/her expressed emotion [23,35]. Discourse information has been combined with acoustic correlates in order to improve the recognition performance of emotion recognition systems [8,3]. In [69], the following speech-acts are used for labeling the user response: rejection, repeat, rephrase, ask-start over, and none of the above. The speech data in this study was obtained from real users engaged in spoken dialog with a machine agent over the telephone using a commercially developed call center application. The main focus of this study was on detecting negative emotions (anger and frustration) versus nonnegative emotions; e.g. neutral and happy. As expected, there is a strong correlation between the speech-act of rejection and the negative emotions. In that work, acoustic, linguistic and discourse information are combined together for recognizing emotions. Linear discriminant classifier (LDC) was used for classification with both linguistic and discourse information. For acoustic information, both the LDC and the k-NN classifier were used. The increase in the classification accuracy due to combining the discourse information with the acoustic information was in the range from 1.4% to 6.75% for male data and 0.75% to 3.96% for female data.

The above information sources can be combined by a variety of ways. The most straightforward way is to combine all measurements output by these sources into one long feature vector [3]. However, as mentioned earlier, having features vectors with high dimensionality is not desirable. Another method is to implement three classifiers, one for each information source, and combine their output decisions using any decision fusion method such as

[1] It is also possible to use a ready-made language model.

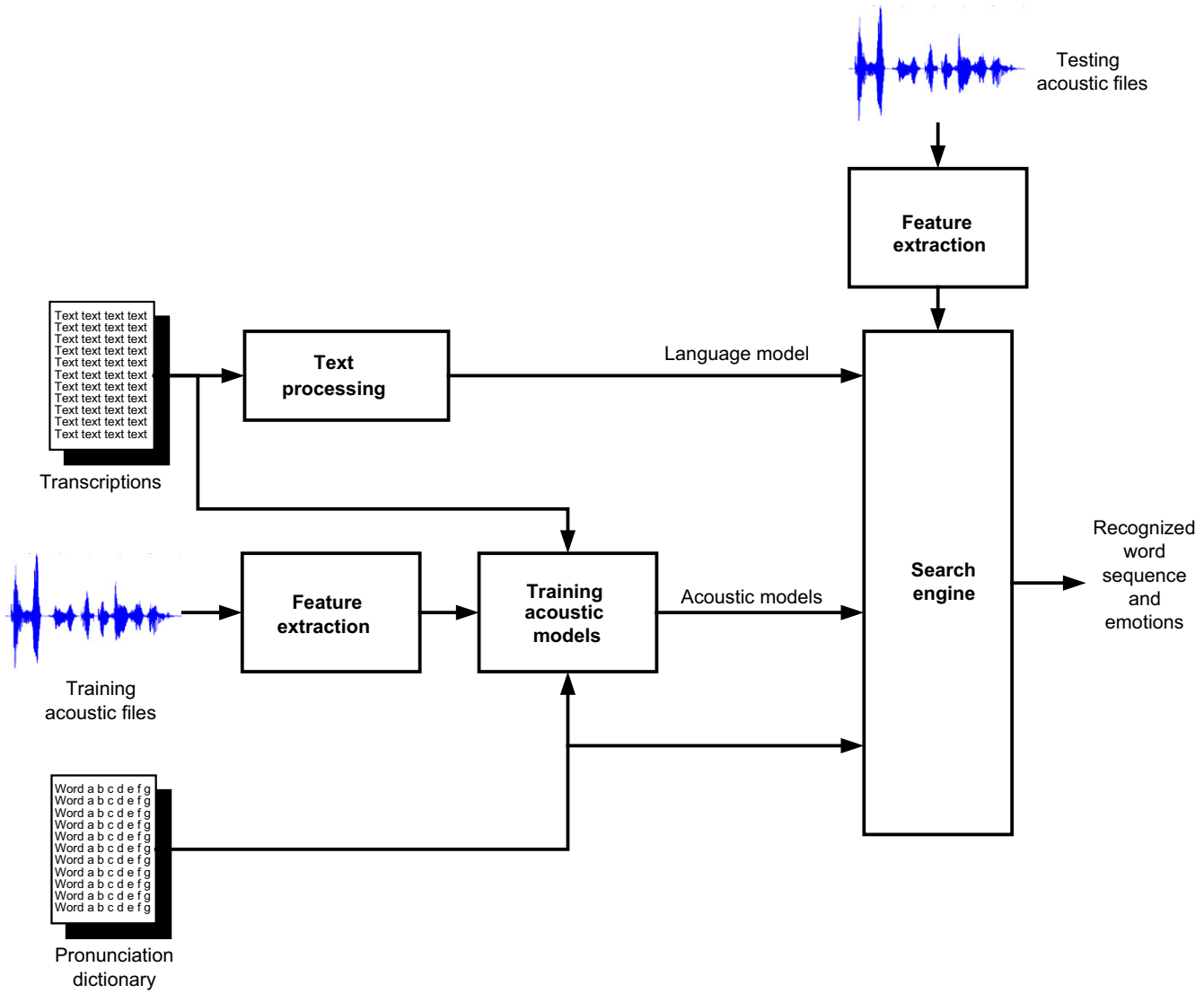

Fig. 2. The architecture of a speech emotion recognition engine combining acoustic and linguistic information.

bagging [16]. In [69], the final decision is based on the average of the likelihood values of all individual classifiers.

3.4.3. *Combining acoustic and video information*

Human facial expressions can be used in detecting emotions. There have been many studies on recognizing emotions based only on video recordings of the facial expressions. [36,97]. According to an experimental study based on human subjective evaluation, De Silva et al. [118] concluded that some emotions are more easily recognized using audio information than using video information and vice versa. Based on this observation, they proposed combining the performances of the audio-based and the video-based systems using any aggregation scheme. In fact, not much research work is done in this area. In this survey, a brief overview of only two studies is given.

The first one is provided in [24]. Regarding speech signal, pitch- and energy-related features such as the minimum and maximum values were first extracted from all the utterances. To analyze the video signal, the Fourier transform of the optical flow vectors for the eye region and the mouth region was computed. This method has shown to be useful in analyzing video sequences [97,77]. The coefficients of the Fourier transform were then used as features for an HMM emotion recognizer. Synchronization was made between the audio and the video signals and all features were pooled in one long vector. The classification scheme was tested using the emotional video corpus developed by De Silva et al. [119]. The corpus contained six emotions: anger, happiness, sadness, surprise, dislike, and fear. The overall decision was made using a rule-based classification approach. Unfortunately, no classification accuracy was reported in this study.

In the other study [45], there were two classifiers: one for the video part and the other for the audio part. The emotional database was locally recorded and contained the basic six archetypal emotions. Features were extracted from the video data using multi-resolution analysis based on the discrete wavelet transform. The dimensionality of the obtained wavelet coefficients vectors was reduced using a combination of the PCA and LDA techniques. In the training phase, a codebook was constructed based on the feature vectors for each emotion. In the testing phase, the extracted features were compared to the reference vectors in each codebook and a membership value was returned. The same was repeated for the audio data. The two obtained membership values for each emotion were combined using the maximum rule. The fusion algorithm was applied to a locally recorded database which contained the following

emotions: happiness, sadness, anger, surprise, fear, and dislike. Speaker-dependent classification was mainly considered in this study. When only acoustic features were used, the recognition accuracies ranged from 57% to 93.3% for male speakers and 68% to 93.3% for female speakers. The facial emotion recognition rates ranged from 65% to 89.4% for male subjects and from 60% to 88.8% for female subjects when the PCA method was used for feature extraction. When the LDA method was used for feature extraction, the accuracies ranged from 70% to 90% for male subjects and from 64.4% to 95% for female subjects. When both acoustic and facial information sources were combined, the recognition accuracies were 98.3% for female speakers and 95% for male speakers.

Finally, it should be mentioned that though the combination of audio and video information seems to be powerful in detecting emotions, the application of such a scheme may not be feasible. Video data may not be available for some applications such as automated dialog systems.

4. Classification schemes

A speech emotion recognition system consists of two stages: (1) a front-end processing unit that extracts the appropriate features from the available (speech) data, and (2) a classifier that decides the underlying emotion of the speech utterance. In fact, most current research in speech emotion recognition has focused on this step since it represents the interface between the problem domain and the classification techniques. On the other hand, traditional classifiers have been used in almost all proposed speech emotion recognition systems.

Various types of classifiers have been used for the task of speech emotion recognition HMM, GMM, SVM, artificial neural networks (ANN), k-NN and many others. In fact, there has been no agreement on which classifier is the most suitable for emotion classification. It seems also that each classifier has its own advantages and limitations. In order to combine the merits of several classifiers, aggregating a group of classifiers has also been recently employed [113,84]. Based on several studies [94,21,72,97,115,43,138,129], we can conclude that HMM is the most used classifier in emotion classification probably because it is widely used in almost all speech applications. The objective of this section is to give an overview of various classifiers used in speech emotion recognition and to discuss the limitation of each one of them. The focus will be on statistical classifiers because they are the most widely used in the context of speech emotion recognition. The classifiers are mentioned according to their relevance in the literature of speech emotion recognition. Multiple classifier systems are also discussed in this section.

In the statistical approach to pattern recognition, each class is modelled by a probability distribution based on the available training data. Statistical classifiers have been used in many speech recognition applications. While HMM is the most widely used classifier in the task of automatic speech recognition (ASR), GMM is considered the state-of-the-art classifier for speaker identification and verification [106].

HMM and GMM generally have many interesting properties such as the ease of implementation and their solid mathematical basis. However, compared to simple parametric classifiers such as LDC and quadratic discriminant analysis (QDC), they have some minor drawbacks compared such as the need of a proper initialization for the model parameters before training and the long training time often associated with them [10].

4.1. Hidden Markov model

The HMM classifier has been extensively used in speech applications such as isolated word recognition and speech segmentation because it is physically related to the production mechanism of speech signal [102]. The HMM is a doubly stochastic process which consists of a first-order Markov chain whose states are *hidden* from the observer. Associated with each state is a random process which generates the observation sequence. Thus, the hidden states of the model capture the temporal structure of the data. Mathematically, for modeling a sequence of observable data vectors, $\mathbf{x}_1,\ldots,\mathbf{x}_T$, by an HMM, we assume the existence of a hidden Markov chain responsible for generating this observable data sequence. Let K be the number of states, π_i, $i=1,\ldots,K$ be the initial state probabilities for the hidden Markov chain, and a_{ij}, $i=1,\ldots,K, j=1,\ldots,K$ be the transition probability from state i to state j. Usually, the HMM parameters are estimated based on the ML principle. Assuming the true state sequence is $s_1,\ldots,s_T$, the likelihood of the observable data is given by

$$p(\mathbf{x}_1,s_1\ldots,\mathbf{x}_T,s_T)=\pi_{s_1}b_{s_1}(\mathbf{x}_1)a_{s_1,s_2}b_{s_2}(\mathbf{x}_2)\ldots a_{s_{T-1},s_T}b_{s_T}(\mathbf{x}_T)$$
$$=\pi_{s_1}b_{s_1}(\mathbf{x}_1)\prod_{t=2}^{T}a_{s_{t-1},s_t}b_{s_t}(\mathbf{x}_t), \quad (5)$$

where

$$b_i(\mathbf{x}_t)\equiv P(\mathbf{x}|s_t=i)$$

is the observation density of the ith state. This density can be either discrete for discrete HMM or a mixture of Gaussian densities for continuous HMM. Since the true state sequence is not typically known, we have to sum over all possible state sequences to find the likelihood of a given data sequence, i.e.

$$p(\mathbf{x}_1,\ldots,\mathbf{x}_T)=\sum_{s_1,\ldots,s_T}\left(\pi_{s_1}b_{s_1}(\mathbf{x}_1)\prod_{t=2}^{T}a_{s_{t-1},s_t}b_{s_t}(\mathbf{x}_t)\right). \quad (6)$$

Fortunately, very efficient algorithms have been proposed for the calculation of the likelihood function in a time of order $\mathcal{O}(KT)$ such as the forward recursion and the backward recursion algorithms (for details about these algorithms, the reader is referred to [102,39]). In the training phase, the HMM parameters are determined as those maximizing the likelihood of (6). This is commonly achieved using the expectation maximization (EM) algorithm [32].

There are many design issues regarding the structure and the training of the HMM classifier. The topology of the HMM may be a left-to-right topology [115] as in most speech recognition applications or a fully connected topology [94]. The assumption of left-to-right topology explicitly models advance in time. However, this assumption may not be valid in the case of speech emotion recognition since, in this case, the HMM states correspond to emotional cues such as pauses. For example, if the pause is associated with the emotion of sadness, there is no definite time instant of this state; the pause may occur at the beginning, the middle, or at the end of the utterance. Thus, any state should be reachable from any other state and a fully connected HMM may be more suitable. Another distinction between ASR and emotion recognition is that the HMM states in the former are aligned with a small number of acoustic features which correspond to small speech units such as phonemes or syllables. On the other hand, prosodic acoustic features associated with emotions only make sense with larger time units spanning at least a word [12]. Other design issues of the HMM classifier include determining the optimal number of states, the type of the observations (discrete versus continuous) and the optimal number of observation symbols (also called codebook size [102]) in case of using discrete HMM or the optimum number of Gaussian components in case of using continuous HMM.

Generally, HMM provides classification accuracies for speech emotion recognition tasks that are comparable to other

well-known classifiers. In [94], an HMM-based system for the classification of the six archetypal emotions was proposed. The LFPC, MFCC, and LPCC were used as a representation of speech signal. A four-state fully connected HMM was built for each emotion and for each speaker. The HMMs were discrete and a codebook of size 64 was constructed for the data of each speaker. Two speech databases were developed by the authors to train and test the HMM classifier: Burmese and Mandarin. Four hundred and thirty-two out of 720 utterances were used for training while the other for testing. The best average rates were 78.5% and 75.5% for the Burmese and Mandarin databases, respectively, while the human classification accuracy was 65.8%. That is, their proposed speech emotion recognition system performed better than human for those particular databases. However, this result cannot be generalized unless a more comprehensive study involving more than database is performed.

HMMs are used in many other studies such as [68,73]. In the former study, the recognition accuracy was 70.1% for 4-class style classification of utterances from the text-independent SUSAS database. In [73], two systems were proposed: the first was an ordinary system in which each emotion was modelled by a continuous HMM system with 12 Gaussian mixtures for each state. In the second system, a three-state continuous HMM was built for each phoneme class. There were 46 phonemes, which were grouped into five classes: vowel, glide, nasal, stop, and fricative sound. Each state was modeled by 16 Gaussian components. The TIMIT speech database was used to train the HMM for each phoneme-class. The evaluation was performed using utterances of another locally recorded emotional speech database which contained the emotions of anger, happiness, neutral, and sadness. Each utterance was segmented to the phoneme level and the phoneme sequence was reported. For each testing utterance, a global HMM was built for this utterance, which was composed of phoneme-class HMMs concatenated in the same order as the corresponding phoneme sequence. The start and end frame numbers of each segment were determined using the Viterbi algorithm. This procedure was repeated for each emotion and the ML criterion was used to determine the expressed emotion. Applying this scheme on a locally recorded speech database containing 704 training utterances and 176 testing utterances, the obtained overall accuracy using the phoneme-class dependent HMM was 76.12% versus 55.68% for SVM using the prosodic features and 64.77% for generic emotional HMM. Based on the obtained results, the authors claimed that phoneme-based modeling provided better discrimination between emotions. This may be true since there are variations across emotional states in the spectral features at the phoneme level, especially vowel sounds [75].

4.2. Gaussian mixture models

Gaussian mixture model is a probabilistic model for density estimation using a convex combination of multi-variate normal densities [133]. It can be considered as a special continuous HMM which contains only one state [107]. GMMs are very efficient in modeling multi-modal distributions [10] and their training and testing requirements are much less than the requirements of a general continuous HMM. Therefore, GMMs are more appropriate for speech emotion recognition when only global features are to be extracted from the training utterances. However, GMMs cannot model temporal structure of the training data since all the training and testing equations are based on the assumption that all vectors are independent. Similar to many other classifiers, determining the optimum number of Gaussian components is an important but difficult problem [107]. The most common way to determine the optimal number of Gaussian components is through model order section criteria such as classification error with respect to a cross validation set, minimum description length (MDL) [108], Akaike information criterion (AIC) [1], and kurtosis-based goodness-of-fit (GOF) measures [37,132]. Recently, a greedy version of the EM algorithm has been developed such that both the model parameters and the model order are estimated simultaneously [133].

In [15], a GMM classifier was used with the KISMET infant-directed speech database, which contains 726 utterances. The emotions encountered were approval, attention, prohibition, soothing, and neutral. A kurtosis-based model selection criterion was used to determine the optimum number of Gaussian components for each model [132]. Due to the limited number of available utterances, a 100-fold cross validation was used to assess the classification performance. The SFS feature selection technique was used to select the best features from a set containing pitch-related and energy-related features. A maximum accuracy of 78.77% accuracy was achieved when the best five features are used. Using a hierarchical sequential classification scheme, the classification accuracy was increased to 81.94%.

The GMM is also used with some other databases such as the BabyEars emotional speech database [120]. This database contains 509 utterances: 212 utterances for the approval emotion, 149 for the attention emotion, and 148 for the prohibition emotion. The cross validation error was measured for a wide range of GMM orders (from 1 to 100). The best average performance obtained was about 75% (speaker-independent classification), which corresponded to a model order of 10. A similar result was obtained with the FERMUS III database [112], which contained a total of 5250 samples for the basic archetypal emotions plus the neutral emotion. Sixteen-component GMMs were used to model each emotion. The average classification accuracy was 74.83% for speaker-independent recognition and 89.12% for speaker-dependent recognition. These results were based on threefold cross validation.

In order to model the temporal structure of the data, the GMM was integrated with the vector autoregressive (VAR) process resulting in what is called Gaussian mixture vector autoregressive model (GMVAR) [6]. The GMVAR model was applied to the Berlin emotional speech database [18] which contained the anger, fear, happiness, boredom, sadness, disgust, and neutral emotions. The disgust emotion was discarded because of the small number of utterances. The GMVAR provided a classification accuracy of 76% versus 71% for the hidden Markov model, 67% for the *k*-nearest neighbors, and 55% for feed-forward neural networks. All the classification accuracies were based on fivefold cross validation where speaker information were not considered in the split of data into training and testing sets; i.e. the classification was speaker dependent. In addition, the GMVAR model provided a 90% accuracy of classification between high-arousal emotions, low-arousal emotions, and the neutral emotion versus 86.00% for the HMM technique.

4.3. Neural networks

Another common classifier, used for many pattern recognition applications is the artificial neural network (ANN). ANNs have some advantages over GMM and HMM. They are known to be more effective in modeling nonlinear mappings. Also, their classification performance is usually better than HMM and GMM when the number of training examples is relatively low. Almost all ANNs can be categorized into three main basic types: MLP, recurrent neural networks (RNN), and radial basis functions (RBF) networks [10]. The latter is rarely used in speech emotion recognition.

MLP neural networks are relatively common in speech emotion recognition. The reason for that may be the ease of implementation and the well-defined training algorithm once the structure of ANN is completely specified. However, ANN classifiers in general have many design parameters, e.g. the form of the neuron activation function, the number of the hidden layers and the number of

neuron in each layer, which are usually set in an ad hoc manner. In fact, the performance of ANN heavily depends on these parameters. Therefore, in some speech emotion recognition systems, more than one ANN is used [93]. An appropriate aggregation scheme is used to combine the outputs of the individuals ANN classifiers.

The classification accuracy of ANN is fairly low compared to other classifiers. In [93], the main objective was to classify the following eight emotions: joy, teasing, fear, sadness, disgust, anger, surprise, and neutral from a locally recorded emotional speech database. The basic classifier was a One-Class-in-One Neural Network (OCON) [87], which consists of eight MLP sub-neural networks and a decision logic control. Each sub-neural network contained two hidden layers in addition to the input and the output layers. The output layer contained only one neuron whose output was an analog value from 0 to 1. Each sub-neural network was trained to recognize one of the eight emotions. In the testing phase, the output of each ANN specified how likely the input speech vectors were produced by a certain emotion. The decision logic control generated a single hypothesis based on the outputs of the eight sub-neural networks. This scheme was applied to a locally recorded speech database, which contained the recordings of 100 speakers. Each speaker uttered 100 words eight times, one for each of the above mentioned emotions. The best classification accuracy was only 52.87%, obtained by training on the utterances of 30 speakers and testing on the remaining utterances; i.e. the classification task was speaker independent. Similar classification accuracies were obtained in [53] with All-Class-in-One neural network architecture. Four topologies were tried in that work. In all of them, the neural network had only one hidden layer which contained 26 neurons. The input layer had either 7 or 8 neurons and the output layer had either 14 or 26 neurons. The best achieved classification accuracy in this work was 51.19%. However, the classification models were speaker dependent.

A better result is found in [99]. In this study, three ANN configurations were applied. The first one was an ordinary two-layer MLP classifier. The speech database was also locally recorded and contained 700 utterances for the following emotions: happiness, anger, sadness, fear, and normal. A subset of the data containing 369 utterances is selected based on subjects' decisions and is randomly split into training (70% of the utterances) and testing (30%) subsets. The average classification accuracy was about 65%. The average classification accuracy was 70% for the second configuration in which the bootstrap aggregation (bagging) scheme was employed. Bagging scheme is a method for generating multiple versions of the classifier and using them to get an aggregated classifier with higher classification accuracy [16]. Finally, an average classification accuracy of 63% was achieved in the third configuration which is very similar to that described in the previous system. The superiority in performance of this study to the other two studies discussed is attributed to the use of different emotional corpus in each study.

4.4. Support vector machine

An important example of the general discriminant classifiers is the support vector machine [34]. SVM classifiers are mainly based on the use of kernel functions to nonlinearly map the original features to a high-dimensional space where data can be well classified using a linear classifier. SVM classifiers are widely used in many pattern recognition applications and shown to outperform other well-known classifiers [70]. They have some advantages over GMM and HMM including the global optimality of the training algorithm [17], and the existence of excellent data-dependent generalization bounds [30]. However, their treatment of nonseparable cases is somewhat heuristic. In fact, there is no systematic way to choose the kernel functions, and hence, separability of the transformed features is not guaranteed. In fact, in many pattern recognition applications including speech emotion recognition, it is not advised to have a perfect separation of the training data so as to avoid over-fitting.

SVM classifiers are also used extensively for the problem of speech emotion recognition in many studies [116,73,68,101]. The performances of almost all of them are similar, and hence, only the first one will be briefly described. In this study, three approaches are investigated in order to extend the basic SVM binary classification to the multi-class case. In the first two approaches, an SVM classifier is used to model each emotion and is trained against all other emotions. In the first approach, the decision is made for the class with highest distance to other classes. In the second approach, the SVM output distances are fed to a 3-layer MLP classifier that produces the final output decision. The third approach followed a hierarchical classification scheme which is described in Section 4.5. The three systems were tested using utterances from the FERMUS III corpus [112]. For speaker-independent classification, the classification accuracies are 76.12%, 75.45%, and 81.29% for the first, the second, and the third approaches, respectively. For speaker-dependent classification, the classification accuracies are 92.95%, 88.7%, and 90.95% for the first, the second, and the third approaches, respectively.

There are many other classifiers that have been applied in many other studies to the problem of speech emotion recognition such as *k*-NN classifiers [116], fuzzy classifiers [105], and decision trees [101]. However, the above-mentioned classifiers, especially the GMM and the HMM, are the most used ones on this task. Moreover, the performance of many of them is not significantly different from the above mentioned classification techniques. Table 2 compares the performance of popular classifiers, employed for the task of speech emotion recognition. One might conclude that the GMM achieves the best compromise between the classification performance and the computational requirements required for training and testing. However, we should be cautious that different emotional corpora with different emotion inventories were used in those individual studies. Moreover, some of those corpora are locally recorded and inaccessible to other researchers. Therefore, such a conclusion cannot be established without performing more comprehensive experiments that employ many accessible corpora for comparing the performance of different classifiers.

4.5. Multiple classifier systems

As an alternative to highly complex classifiers that may require large computational requirement for training, multiple classifier

Table 2
Classification performance of popular classifiers, employed for the task of speech emotion recognition.

Classifier	HMM	GMM	ANN	SVM
Average classification accuracy	75.5–78.5% [94,115]	74.83–81.94% [15,120]	51.19–52.82% [93,53] 63–70% [99]	75.45–81.29% [116]
Average training time	Small	Smallest	Back-propagation: large	Large
Sensitivity to model initialization	Sensitive	Sensitive	Sensitive	Insensitive

systems (MCS) have been proposed recently for the task of speech emotion recognition [113,84]. There are three approaches for combining classifiers [67,84]: hierarchical, serial, and parallel. In the hierarchical approach, classifiers are arranged in a tree structure where the set of candidate classes becomes smaller as we go in depth in the tree. At the leave-node classifiers, only one class remains after decision. In the serial approach, classifiers are placed in a queue where each classifier reduces the number of candidate classes for the next classifier [88,139]. In the parallel approach, all classifiers work independently and a decision fusion algorithm is applied to their outputs [66].

The hierarchical approach was applied in [83] for classifying utterances from the Berlin emotional database [18] where the main goal was to improve speaker-independent emotion classification. The following emotions are selected for classification: anger, happiness, sadness, boredom, anxiety, and neutral. The hierarchical classification system was motivated by the psychological study of emotions in [110] in which emotions are represented in three dimensions: activation (arousal), potency (power), and evaluation (pleasure). Therefore, 2-stage and 3-stage hierarchical classification systems were proposed in [83]. The naive Bayesian classifier [34] was used for all classifications. Both systems are shown in Fig. 3. Prosody features included statistics of pitch, energy, duration, articulation, and zero-crossing rate. Voice quality features were calculated as parameters of the excitation spectrum, called spectral gradients [121]. The 2-stage system provided a classification accuracy of 83.5% which is about 9% more than that obtained by the same authors in a previous study using the same voice quality features [82]. For 3-stage classification, the classification accuracy is further increased to 88.8%. In the two studies, classification accuracies are based on 10-fold cross validation but the validation data vectors were used for both feature selection (they used Sequential Floating Forward Search (SFFS) algorithm) and testing.

All the three approaches for combining classifiers were applied to speech emotion recognition in [84]. The authors applied the same experimental setup as in the previous study. When the validation vectors were used for both feature selection and testing, the classification accuracies for the hierarchical, the serial, and the parallel approaches for classifier combination were 88.6%, 96.5%, and 92.6%, respectively, versus 74.6%. When the validation and test data sets are different, the classification accuracies reduce considerably to 58.6%, 59.7%, 61.8%, and 70.1% for single classifier, the hierarchical approach, the serial approach, and the parallel approach for combining classifiers, respectively.

5. Conclusions

In this paper, a survey of current research work in speech emotion recognition system has been given. Three important issues have been studied: the features used to characterize different emotions, the classification techniques used in previous research, and the important design criteria of emotional speech databases. There are several conclusions that can be drawn from this study.

The first one is that while high classification accuracies have been obtained for classification between high-arousal and low-arousal emotions, N-way classification is still challenging. Moreover, the performance of current stress detectors still needs significant improvement. The average classification accuracy of speaker-independent speech emotion recognition systems is less than 80% in most of the proposed techniques. In some cases, such as [93], it is as low as 50%. For speaker-dependent classification, the recognition accuracy exceeded 90% only in few studies [116,101,98]. Many classifiers have been tried for speech emotion recognition such as the HMM, the GMM, the ANN, and the SVM. However, it is hard to decide which classifier performs best for this task because different emotional corpora with different experimental setups were applied.

Most of the current body of research focuses on studying many speech features and their relations to the emotional content of the speech utterance. New features have also been developed such as the TEO-based features. There are also attempts to employ different feature selection techniques in order to find the best features for this task. However, the conclusions obtained from different studies are not consistent. The main reason may be attributed to the fact that only one emotional speech database is investigated in each study.

Most of the existing databases are not perfect for evaluating the performance of a speech emotion recognizer. In many databases, it is difficult even for human subjects to determine the emotion of some recorded utterances; e.g. the human recognition accuracy was 67% for DED [38], 80% for Berlin [18], and 65% in [94]. There are some other problems for some databases such as the low quality of the recorded utterances, the small number of available utterances, and the unavailability of phonetic transcriptions. Therefore, it is likely that some of the conclusions established in some studies cannot be generalized to other databases. To address this problem, more cooperation across research institutes in developing benchmark emotional speech databases is necessary.

In order to improve the performance of current speech emotion recognition systems, the following possible extensions are proposed. The first extension relies on the fact that speaker-dependent classification is generally easier than speaker-independent classification. At the same time, there exist speaker identification techniques with high recognition performance such as the GMM-based text-independent speaker identification system proposed by Reynolds [107]. Thus, a speaker-independent emotion recognition system may be implemented as a combination of a speaker identification system followed by a speaker-dependent emotion recognition system.

It is also noted that the majority of the existing classification techniques do not model the temporal structure of the training data. The only exception may be the HMM in which time

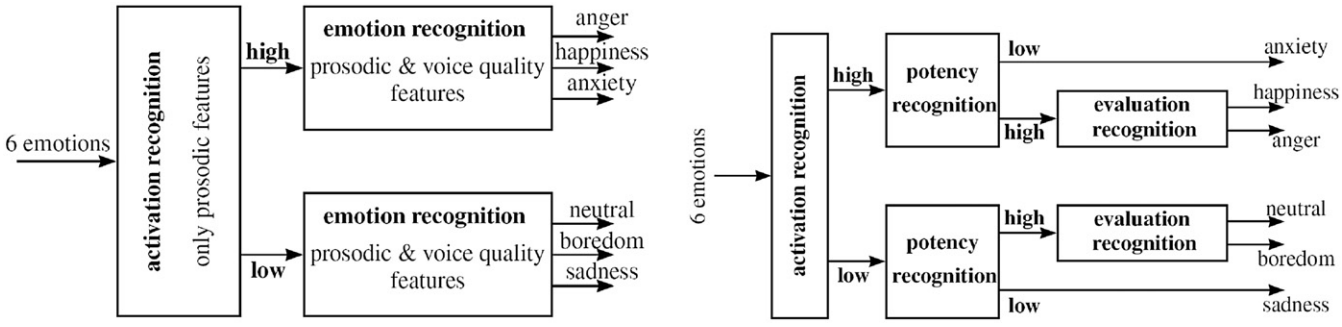

Fig. 3. 2-stage and 3-stage hierarchical classification of emotions by Lugger and Yang [83].

dependency may be modelled using its states. However, all the Baum–Welch re-estimation formulae are based on the assumption that all the feature vectors are statistically independent [102]. This assumption is invalid in practice. It is sought that direct modeling of the dependency between feature vectors, e.g. through the use of autoregressive models, may provide an improvement in the classification performance. Potential discriminative sequential classifiers that do not assume statistical independence between feature vectors include conditional random fields (CRF) [134] and switching linear dynamic system (SLDS) [89].

Finally, there are only few studies that considered applying multiple classifier systems (MCS) to speech emotion recognition [84,113]. We believe that this research direction has to be further explored. In fact, MCS is now a well-established area in pattern recognition [66,67,127,126] and there are many aggregation techniques that have not been applied to speech emotion recognition such as Adaboost.M1 [42] and dynamic classifier selection (DCS) [52].

References

[1] H. Akaike, A new look at the statistical model identification, IEEE Trans. Autom. Control 19 (6) (1974) 716–723.
[2] N. Amir, S. Ron, N. Laor, Analysis of an emotional speech corpus in Hebrew based on objective criteria, in: SpeechEmotion-2000, 2000, pp. 29–33.
[3] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, A. Stolcke, Prosody-based automatic detection of annoyance and frustration in human–computer dialog, in: Proceedings of the ICSLP 2002, 2002, pp. 2037–2040.
[4] B.S. Atal, Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification, J. Acoust. Soc. Am. 55 (6) (1974) 1304–1312.
[5] T. Athanaselis, S. Bakamidis, I. Dologlou, R. Cowie, E. Douglas-Cowie, C. Cox, Asr for emotional speech: clarifying the issues and enhancing the performance, Neural Networks 18 (2005) 437–444.
[6] M.M.H. El Ayadi, M.S. Kamel, F. Karray, Speech emotion recognition using Gaussian mixture vector autoregressive models, in: ICASSP 2007, vol. 4, 2007, pp. 957–960.
[7] R. Banse, K. Scherer, Acoustic profiles in vocal emotion expression, J. Pers. Soc. Psychol. 70 (3) (1996) 614–636.
[8] A. Batliner, K. Fischer, R. Huber, J. Spiker, E. Noth, Desperately seeking emotions: actors, wizards and human beings, in: Proceedings of the ISCA Workshop Speech Emotion, 2000, pp. 195–200.
[9] S. Beeke, R. Wilkinson, J. Maxim, Prosody as a compensatory strategy in the conversations of people with agrammatism, Clin. Linguist. Phonetics 23 (2) (2009) 133–155.
[10] C.M. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, 1995.
[11] M. Borchert, A. Dusterhoft, Emotions in speech—experiments with prosody and quality features in speech for use in categorical and dimensional emotion recognition environments, in: Proceedings of 2005 IEEE International Conference on Natural Language Processing and Knowledge Engineering, IEEE NLP-KE'05 2005, 2005, pp. 147–151.
[12] L. Bosch, Emotions, speech and the asr framework, Speech Commun. 40 (2003) 213–225.
[13] S. Bou-Ghazale, J. Hansen, A comparative study of traditional and newly proposed features for recognition of speech under stress, IEEE Trans. Speech Audio Process. 8 (4) (2000) 429–442.
[14] R. Le Bouquin, Enhancement of noisy speech signals: application to mobile radio communications, Speech Commun. 18 (1) (1996) 3–19.
[15] C. Breazeal, L. Aryananda, Recognition of affective communicative intent in robot-directed speech, Autonomous Robots 2 (2002) 83–104.
[16] L. Breiman, Bagging predictors, Mach. Learn. 24 (2) (1996) 123–140.
[17] C.J.C. Burges, A tutorial on support vector machines for pattern recognition, Data Mining Knowl. Discovery 2 (2) (1998) 121–167.
[18] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, B. Weiss, A database of German emotional speech, in: Proceedings of the Interspeech 2005, Lissabon, Portugal, 2005, pp. 1517–1520.
[19] C. Busso, S. Lee, S. Narayanan, Analysis of emotionally salient aspects of fundamental frequency for emotion detection, IEEE Trans. Audio Speech Language Process. 17 (4) (2009) 582–596.
[20] J. Cahn, The generation of affect in synthesized speech, J. Am. Voice Input/Output Soc. 8 (1990) 1–19.
[21] D. Caims, J. Hansen, Nonlinear analysis and detection of speech under stressed conditions, J. Acoust. Soc. Am. 96 (1994) 3392–3400.
[22] W. Campbell, Databases of emotional speech, in: Proceedings of the ISCA (International Speech Communication and Association) ITRW on Speech and Emotion, 2000, pp. 34–38.
[23] C. Chen, M. You, M. Song, J. Bu, J. Liu, An enhanced speech emotion recognition system based on discourse information, in: Lecture Notes in Computer Science—I (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 3991, 2006, pp. 449–456, cited by (since 1996) 1.
[24] L. Chen, T. Huang, T. Miyasato, R. Nakatsu, Multimodal human emotion/expression recognition, in: Proceedings of the IEEE Automatic Face and Gesture Recognition, 1998, pp. 366–371.
[25] Z. Chuang, C. Wu, Emotion recognition using acoustic features and textual content, Multimedia and Expo, 2004. IEEE International Conference on ICME '04, vol. 1, 2004, pp. 53–56.
[26] R. Cohen, A computational theory of the function of clue words in argument understanding, in: ACL-22: Proceedings of the 10th International Conference on Computational Linguistics and 22nd Annual Meeting on Association for Computational Linguistics, 1984, pp. 251–258.
[27] R. Cowie, R.R. Cornelius, Describing the emotional states that are expressed in speech, Speech Commun. 40 (1–2) (2003) 5–32.
[28] R. Cowie, E. Douglas-Cowie, Automatic statistical analysis of the signal and prosodic signs of emotion in speech, in: Proceedings, Fourth International Conference on Spoken Language, 1996. ICSLP 96. vol. 3, 1996, pp. 1989–1992.
[29] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, S. Kollias, W. Fellenz, J. Taylor, Emotion recognition in human–computer interaction, IEEE Signal Process. Mag. 18 (2001) 32–80.
[30] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines, Cambridge University Press, 2000.
[31] J.R. Davitz, The Communication of Emotional Meaning, McGraw-Hill, New York, 1964.
[32] A. Dempster, N. Laird, D. Rubin, Maximum likelihood from incomplete data via the em algorithm, J. R. Stat. Soc. 39 (1977) 1–38.
[33] L. Devillers, L. Lamel, Emotion detection in task-oriented dialogs, in: Proceedings of the International Conference on Multimedia and Expo 2003, 2003, pp. 549–552.
[34] R. Duda, P. Hart, D. Stork, Pattern Recognition, John Wiley and Sons, 2001.
[35] D. Edwards, Emotion discourse, Culture Psychol. 5 (3) (1999) 271–291.
[36] P. Ekman, Emotion in the Human Face, Cambridge University Press, Cambridge, 1982.
[37] M. Abu El-Yazeed, M. El Gamal, M. El Ayadi, On the determination of optimal model order for gmm-based text-independent speaker identification, EURASIP J. Appl. Signal Process. 8 (2004) 1078–1087.
[38] I. Engberg, A. Hansen, Documentation of the Danish emotional speech database des 〈http://cpk.auc.dk/tb/speech/Emotions/〉, 1996.
[39] Y. Ephraim, N. Merhav, Hidden Markov processes, IEEE Trans. Inf. Theory 48 (6) (2002) 1518–1569.
[40] R. Fernandez, A computational model for the automatic recognition of affect in speech, Ph.D. Thesis, Massachusetts Institute of Technology, February 2004.
[41] D.J. France, R.G. Shiavi, S. Silverman, M. Silverman, M. Wilkes, Acoustical properties of speech as indicators of depression and suicidal risk, IEEE Trans. Biomedical Eng. 47 (7) (2000) 829–837.
[42] Y. Freund, R.E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, J. Comput. Syst. Sci. 55 (1) (1997) 119–139 cited by (since 1996) 1695.
[43] L. Fu, X. Mao, L. Chen, Speaker independent emotion recognition based on svm/hmms fusion system, in: International Conference on Audio, Language and Image Processing, 2008. ICALIP 2008, pp. 61–65.
[44] M. Gelfer, D. Fendel, Comparisons of jitter, shimmer, and signal-to-noise ratio from directly digitized versus taped voice samples, J. Voice 9 (4) (1995) 378–382.
[45] H. Go, K. Kwak, D. Lee, M. Chun, Emotion recognition from the facial image and speech signal, in: Proceedings of the IEEE SICE 2003, vol. 3, 2003, pp. 2890–2895.
[46] C. Gobl, A.N. Chasaide, The role of voice quality in communicating emotion, mood and attitude, Speech Commun. 40 (1–2) (2003) 189–212.
[47] A. Gorin, On automated language acquisition, J. Acoust. Soc. Am. 97 (1995) 3441–3461.
[48] B.J. Grosz, C.L. Sidner, Attention, intentions, and the structure of discourse, Comput. Linguist. 12 (3) (1986) 175–204.
[49] J. Hansen, D. Cairns, Icarus: source generator based real-time recognition of speech in noisy stressful and Lombard effect environments, Speech Commun. 16 (4) (1995) 391–422.
[50] J. Hernando, C. Nadeu, Linear prediction of the one-sided autocorrelation sequence for noisy speech recognition, IEEE Trans. Speech Audio Process. 5 (1) (1997) 80–84.
[51] K. Hirose, H. Fujisaki, M. Yamaguchi, Synthesis by rule of voice fundamental frequency contours of spoken Japanese from linguistic information, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '84, vol. 9, 1984, pp. 597–600.
[52] T. Ho, J. Hull, S.N. Srihari, Decision combination in multiple classifier systems, IEEE Trans. Pattern Anal. Mach. Intell. 16 (1) (1994) 66–75.
[53] V. Hozjan, Z. Kacic, Context-independent multilingual emotion recognition from speech signal, Int. J. Speech Technol. 6 (2003) 311–320.
[54] V. Hozjan, Z. Moreno, A. Bonafonte, A. Nogueiras, Interface databases: design and collection of a multilingual emotional speech database, in: Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC'02) Las Palmas de Gran Canaria, Spain, 2002, pp. 2019–2023.
[55] H. Hu, M. Xu, W. Wu, Dimensions of emotional meaning in speech, in: Proceedings of the ISCA ITRW on Speech and Emotion, 2000, pp. 25–28.

[56] H. Hu, M. Xu, W. Wu, Gmm supervector based svm with spectral features for speech emotion recognition, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 2007. ICASSP 2007, vol. 4, 2007, pp. IV 413–IV 416.
[57] H. Hu, M.-X. Xu, W. Wu, Fusion of global statistical and segmental spectral features for speech emotion recognition, in: International Speech Communication Association—8th Annual Conference of the International Speech Communication Association, Interspeech 2007, vol. 2, 2007, pp. 1013–1016.
[58] A.K. Jain, R.P.W. Duin, J. Mao, Statistical pattern recognition: a review, IEEE Trans. Pattern Anal. Mach. Intell. 22 (1) (2000) 4–37.
[59] T. Johnstone, C.M. Van Reekum, K. Hird, K. Kirsner, K.R. Scherer, Affective speech elicited with a computer game, Emotion 5 (4) (2005) 513–518 cited by (since 1996) 6.
[60] T. Johnstone, K.R. Scherer, Vocal Communication of Emotion, second ed., Guilford, New York, 2000, pp. 226–235.
[61] J. Deller Jr., J. Proakis, J. Hansen, Discrete Time Processing of Speech Signal, Macmillan, 1993.
[62] P.R. Kleinginna Jr., A.M. Kleinginna, A categorized list of emotion definitions, with suggestions for a consensual definition, Motivation Emotion 5 (4) (1981) 345–379.
[63] J. Kaiser, On a simple algorithm to calculate the 'energy' of the signal, in: ICASSP-90, 1990, pp. 381–384.
[64] L. Kaiser, Communication of affects by single vowels, Synthese 14 (4) (1962) 300–319.
[65] E. Kim, K. Hyun, S. Kim, Y. Kwak, Speech emotion recognition using eigen-fft in clean and noisy environments, in: The 16th IEEE International Symposium on Robot and Human Interactive Communication, 2007, RO-MAN 2007, 2007, pp. 689–694.
[66] L.I. Kuncheva, A theoretical study on six classifier fusion strategies, IEEE Trans. Pattern Anal. Mach. Intell. 24 (2002) 281–286.
[67] L.I. Kuncheva, Combining Pattern Classifiers: Methods and Algorithms, Wiley, 2004.
[68] O. Kwon, K. Chan, J. Hao, T. Lee, Emotion recognition by speech signal, in: EUROSPEECH Geneva, 2003, pp. 125–128.
[69] C. Lee, S. Narayanan, Toward detecting emotions in spoken dialogs, IEEE Trans. Speech Audio Process. 13 (2) (2005) 293–303.
[70] C. Lee, S. Narayanan, R. Pieraccini, Classifying emotions in human–machine spoken dialogs, in: Proceedings of the ICME'02, vol. 1, 2002, pp. 737–740.
[71] C. Lee, S.S. Narayanan, R. Pieraccini, Classifying emotions in human–machine spoken dialogs, in: 2002 IEEE International Conference on Multimedia and Expo, 2002, ICME '02, Proceedings, vol. 1, 2002, pp. 737–740.
[72] C. Lee, R. Pieraccini, Combining acoustic and language information for emotion recognition, in: Proceedings of the ICSLP 2002, 2002, pp. 873–876.
[73] C. Lee, S. Yildrim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, S. Narayanan, Emotion recognition based on phoneme classes, in: Proceedings of ICSLP, 2004, pp. 2193–2196.
[74] L. Leinonen, T. Hiltunen, Expression of emotional-motivational connotations with a one-word utterance, J. Acoust. Soc. Am. 102 (3) (1997) 1853–1863.
[75] L. Leinonen, T. Hiltunen, I. Linnankoski, M. Laakso, Expression of emotional-motivational connotations with a one-word utterance, J. Acoust. Soc. Am. 102 (3) (1997) 1853–1863.
[76] X. Li, J. Tao, M.T. Johnson, J. Soltis, A. Savage, K.M. Leong, J.D. Newman, Stress and emotion classification using jitter and shimmer features, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 2007. ICASSP 2007, vol. 4, April 2007, pp. IV-1081–IV-1084.
[77] J. Lien, T. Kanade, C. Li, Detection, tracking and classification of action units in facial expression, J. Robotics Autonomous Syst. 31 (3) (2002) 131–146.
[78] University of Pennsylvania Linguistic Data Consortium, Emotional prosody speech and transcripts 〈http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002S28〉, July 2002.
[79] J. Liscombe, Prosody and speaker state: paralinguistics, pragmatics, and proficiency, Ph.D. Thesis, Columbia University, 2007.
[80] D.G. Lowe, Object recognition from local scale-invariant features, in: Proceedings of the IEEE International Conference on Computer Vision, vol. 2, 1999, pp. 1150–1157.
[81] M. Lugger, B. Yang, The relevance of voice quality features in speaker independent emotion recognition, in: icassp, vol. 4, 2007, pp. 17–20.
[82] M. Lugger, B. Yang, The relevance of voice quality features in speaker independent emotion recognition, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 2007, ICASSP 2007, vol. 4, April 2007, pp. IV-17–IV-20.
[83] M. Lugger, B. Yang, Psychological motivated multi-stage emotion classification exploiting voice quality features, in: F. Mihelic, J. Zibert (Eds.), Speech Recognition, In-Tech, 2008.
[84] M. Lugger, B. Yang, Combining classifiers with diverse feature sets for robust speaker independent emotion recognition, in: Proceedings of EUSIPCO, 2009.
[85] M. Lugger, B. Yang, W. Wokurek, Robust estimation of voice quality parameters under realworld disturbances, in: 2006 IEEE International Conference on Acoustics, Speech and Signal Processing, 2006, ICASSP 2006 Proceedings, vol. 1, May 2006, pp. I–I.
[86] J. Ma, H. Jin, L. Yang, J. Tsai, in: Ubiquitous Intelligence and Computing: Third International Conference, UIC 2006, Wuhan, China, September 3–6, 2006, Proceedings (Lecture Notes in Computer Science), Springer-Verlag, New York, Inc., Secaucus, NJ, USA, 2006.
[87] J. Markel, A. Gray, Linear Prediction of Speech, Springer-Verlag, 1976.
[88] D. Mashao, M. Skosan, Combining classifier decisions for robust speaker identification, Pattern Recognition 39 (1) (2006) 147–155.
[89] B. Mesot, D. Barber, Switching linear dynamical systems for noise robust speech recognition, IEEE Trans. Audio Speech Language Process. 15 (6) (2007) 1850–1858.
[90] P. Mitra, C. Murthy, S. Pal, Unsupervised feature selection using feature similarity, IEEE Trans. Pattern Anal. Mach. Intell. 24 (3) (2002) 301–312.
[91] D. Morrison, R. Wang, L. De Silva, Ensemble methods for spoken emotion recognition in call-centres, Speech Commun. 49 (2) (2007) 98–112.
[92] I. Murray, J. Arnott, Toward a simulation of emotions in synthetic speech: A review of the literature on human vocal emotion, J. Acoust. Soc. Am. 93 (2) (1993) 1097–1108.
[93] J. Nicholson, K. Takahashi, R. Nakatsu, Emotion recognition in speech using neural networks, Neural Comput. Appl. 9 (2000) 290–296.
[94] T. Nwe, S. Foo, L. De Silva, Speech emotion recognition using hidden Markov models, Speech Commun. 41 (2003) 603–623.
[95] J. O'Connor, G. Arnold, Intonation of Colloquial English, second ed., Longman, London, UK, 1973.
[96] A. Oster, A. Risberg, The identification of the mood of a speaker by hearing impaired listeners, Speech Transmission Lab. Quarterly Progress Status Report 4, Stockholm, 1986, pp. 79–90.
[97] T. Otsuka, J. Ohya, Recognizing multiple persons' facial expressions using hmm based on automatic extraction of significant frames from image sequences, in: Proceedings of the International Conference on Image Processing (ICIP-97), 1997, pp. 546–549.
[98] T.L. Pao, Y.-T. Chen, J.-H. Yeh, W.-Y. Liao, Combining acoustic features for improved emotion recognition in Mandarin speech, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 3784, 2005, pp. 279–285, cited by (since 1996) 1.
[99] V. Petrushin, Emotion recognition in speech signal: experimental study, development and application, in: Proceedings of the ICSLP 2000, 2000, pp. 222–225.
[100] R.W. Picard, E. Vyzas, J. Healey, Toward machine emotional intelligence: analysis of affective physiological state, IEEE Trans. Pattern Anal. Mach. Intell. 23 (10) (2001) 1175–1191.
[101] O. Pierre-Yves, The production and recognition of emotions in speech: features and algorithms, Int. J. Human–Computer Stud. 59 (2003) 157–183.
[102] L. Rabiner, B. Juang, An introduction to hidden Markov models, IEEE ASSP Mag. 3 (1) (1986) 4–16.
[103] L. Rabiner, B. Juang, Fundamentals of Speech Recognition, Prentice Hall, 1993.
[104] L. Rabiner, R. Schafer, Digital Processing of Speech Signals, first ed., Pearson Education, 1978.
[105] A. Razak, R. Komiya, M. Abidin, Comparison between fuzzy and nn method for speech emotion recognition, in: 3rd International Conference on Information Technology and Applications ICITA 2005, vol. 1, 2005, pp. 297–302.
[106] D. Reynolds, T. Quatieri, R. Dunn, Speaker verification using adapted Gaussian mixture models, Digital Signal Process. 10 (2000) 19–41.
[107] D. Reynolds, C. Rose, Robust text-independent speaker identification using Gaussian mixture speaker models, IEEE Trans. Speech Audio Process. 3 (1) (1995) 72–83.
[108] J. Rissanen, Modeling by shortest data description, Automatica 14 (5) (1978) 465–471.
[109] K.R. Scherer, Vocal affect expression. A review and a model for future research, Psychological Bull. 99 (2) (1986) 143–165 cited by (since 1996) 311.
[110] H. Schlosberg, Three dimensions of emotion, Psychological Rev. 61 (2) (1954) 81–88.
[111] M. Schubiger, English intonation: its form and function, Niemeyer, Tubingen, Germany, 1958.
[112] B. Schuller, Towards intuitive speech interaction by the integration of emotional aspects, in: 2002 IEEE International Conference on Systems, Man and Cybernetics, vol. 6, 2002, p. 6.
[113] B. Schuller, M. Lang, G. Rigoll, Robust acoustic speech emotion recognition by ensembles of classifiers, in: Proceedings of the DAGA'05, 31, Deutsche Jahrestagung für Akustik, DEGA, 2005, pp. 329–330.
[114] B. Schuller, S. Reiter, R. Muller, M. Al-Hames, M. Lang, G. Rigoll, Speaker independent speech emotion recognition by ensemble classification, in: IEEE International Conference on Multimedia and Expo, 2005. ICME 2005, 2005, pp. 864–867.
[115] B. Schuller, G. Rigoll, M. Lang, Hidden Markov model-based speech emotion recognition, in: International Conference on Multimedia and Expo (ICME), vol. 1, 2003, pp. 401–404.
[116] B. Schuller, G. Rigoll, M. Lang, Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture, in: Proceedings of the ICASSP 2004, vol. 1, 2004, pp. 577–580.
[117] M.T. Shami, M.S. Kamel, Segment-based approach to the recognition of emotions in speech, in: IEEE International Conference on Multimedia and Expo, 2005. ICME 2005, 2005, 4pp.
[118] L.C. De Silva, T. Miyasato, R. Nakatsu, Facial emotion recognition using multimodal information, in: Proceedings of the IEEE International Conference on Information, Communications and Signal Processing (ICICS'97), 1997, pp. 397–401.
[119] L.C. De Silva, T. Miyasato, R. Nakatsu, Facial emotion recognition using multimodal information, in: Proceedings of 1997 International Conference on

Information, Communications and Signal Processing, 1997, ICICS, vol. 1, September 1997, pp. 397–401.
[120] M. Slaney, G. McRoberts, Babyears: a recognition system for affective vocalizations, Speech Commun. 39 (2003) 367–384.
[121] K. Stevens, H. Hanson, Classification of glottal vibration from acoustic measurements, Vocal Fold Physiol. (1994) 147–170.
[122] R. Sun, E. Moore, J.F. Torres, Investigating glottal parameters for differentiating emotional categories with similar prosodics, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 2009. ICASSP 2009, April 2009, pp. 4509–4512.
[123] J. Tao, Y. Kang, A. Li, Prosody conversion from neutral speech to emotional speech, IEEE Trans. Audio Speech Language Process. 14 (4) (2006) 1145–1154.
[124] H. Teager, Some observations on oral air flow during phonation, IEEE Trans. Acoust. Speech Signal Process. 28 (5) (1990) 599–601.
[125] H. Teager, S. Teager, Evidence for nonlinear production mechanisms in the vocal tract, in: Speech Production and Speech Modelling, Nato Advanced Institute, vol. 55, 1990, pp. 241–261.
[126] A. Tsymbal, M. Pechenizkiy, P. Cunningham, Diversity in search strategies for ensemble feature selection, Inf. Fusion 6 (32) (2005) 146–156.
[127] A. Tsymbal, S. Puuronen, D.W. Patterson, Ensemble feature selection with the simple Bayesian classification, Inf. Fusion 4 (32) (2003) 146–156.
[128] D. Ververidis, C. Kotropoulos, Emotional speech classification using Gaussian mixture models and the sequential floating forward selection algorithm, in: IEEE International Conference on Multimedia and Expo, 2005. ICME 2005, July 2005, pp. 1500–1503.
[129] D. Ververidis, C. Kotropoulos, Emotional speech recognition: resources, features and methods, Speech Commun. 48 (9) (2006) 1162–1181.
[130] D. Ververidis, C. Kotropoulos, I. Pitas, Automatic emotional speech classification, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004, Proceedings, (ICASSP '04), vol. 1, 2004, pp. I-593-6.
[131] A. Viterbi, Error bounds for convolutional codes and an asymptotically optimum decoding algorithm Viterbi, IEEE Trans. Inf. Theory 13 (2) (1967) 260–269.
[132] N. Vlassis, A. Likas, A kurtosis-based dynamic approach to Gaussian mixture modeling, IEEE Trans. Syst. Man Cybern. 29 (4) (1999) 393–399.
[133] N. Vlassis, A. Likas, A greedy em algorithm for Gaussian mixture learning, Neural Process. Lett. 15 (2002) 77–87.
[134] Y. Wang, K.-F. Loe, J.-K. Wu, A dynamic conditional random field model for foreground and shadow segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 28 (2) (2006) 279–289.
[135] C. Williams, K. Stevens, Emotions and speech: some acoustical correlates, J. Acoust. Soc. Am. 52 (4 Pt 2) (1972) 1238–1250.
[136] C. Williams, K. Stevens, Vocal correlates of emotional states, Speech Evaluation in Psychiatry, Grune and Stratton, 1981, pp. 189–220.
[137] I. Witten, E. Frank, Data Mining, Morgan Kauffmann, Los Atlos, CA, 2000.
[138] B.D. Womack, J.H.L. Hansen, N-channel hidden Markov models for combined stressed speech classification and recognition, IEEE Trans. Speech Audio Process. 7 (6) (1999) 668–677.
[139] J. Wu, M.D. Mullin, J.M. Rehg, Linear asymmetric classifier for cascade detectors, in: 22th International Conference on Machine Learning, 2005.
[140] M. You, C. Chen, J. Bu, J. Liu, J. Tao, Getting started with susas: a speech under simulated and actual stress database, in: EUROSPEECH-97, vol. 4, 1997, pp. 1743–1746.
[141] M. You, C. Chen, J. Bu, J. Liu, J. Tao, Emotion recognition from noisy speech, in: IEEE International Conference on Multimedia and Expo, 2006, 2006, pp. 1653–1656l.
[142] M. You, C. Chen, J. Bu, J. Liu, J. Tao, Emotional speech analysis on nonlinear manifold, in: 18th International Conference on Pattern Recognition, 2006. ICPR 2006, vol. 3, 2006, pp. 91–94.
[143] M. You, C. Chen, J. Bu, J. Liu, J. Tao, A hierarchical framework for speech emotion recognition, in: IEEE International Symposium on Industrial Electronics, 2006, vol. 1, 2006, pp. 515–519.
[144] S. Young, Large vocabulary continuous speech recognition, IEEE Signal Process. Mag. 13 (5) (1996) 45–57.
[145] G. Zhou, J. Hansen, J. Kaiser, Nonlinear feature based classification of speech under stress, IEEE Trans. Speech Audio Process. 9 (3) (2001) 201–216.
[146] J. Zhou, G. Wang, Y. Yang, P. Chen, Speech emotion recognition based on rough set and svm, in: 5th IEEE International Conference on Cognitive Informatics, 2006, ICCI 2006, vol. 1, 2006, pp. 53–61.

Moataz M.H. El Ayadi received his B.Sc. degree (Hons) in Electronics and Communication Engineering, Cairo University, in 2000, M.Sc. degree in Engineering Mathematics and Physics, Cairo University, in 2004, and Ph.D. degree in Electrical and Computer Engineering, University of Waterloo, in 2008.

He worked as a postdoctoral research fellow in the Electrical and Computer Engineering Department, University of Toronto, from January 2009 to March 2010. Since April 2010, has been an assistant professor in the Engineering Mathematics and Physics Department, Cairo University.

His research interests include statistical pattern recognition and speech processing. His master work was in enhancing the performance of text independent speaker identification systems that uses Gaussian Mixture Models as the core statistical classifier. The main contribution was in developing a new model order selection technique based on the goodness of fit statistical test. He is expected to follow the same line of research in his Ph.D.

Mohamed S. Kamel received the B.Sc. (Hons) EE (Alexandria University), M.A.Sc. (McMaster University), Ph.D. (University of Toronto).

He joined the University of Waterloo, Canada, in 1985 where he is at present Professor and Director of the Pattern Analysis and Machine Intelligence Laboratory at the Department of Electrical and Computer Engineering and holds a University Research Chair. Professor Kamel held Canada Research Chair in Cooperative Intelligent Systems from 2001 to 2008.

Dr. Kamel's research interests are in Computational Intelligence, Pattern Recognition, Machine Learning and Cooperative Intelligent Systems. He has authored and co-authored over 390 papers in journals and conference proceedings, 11 edited volumes, two patents and numerous technical and industrial project reports. Under his supervision, 81 Ph.D. and M.A.Sc. students have completed their degrees.

He is the Editor-in-Chief of the International Journal of Robotics and Automation, Associate Editor of the IEEE SMC, Part A, Pattern Recognition Letters, Cognitive Neurodynamics journal and Pattern Recognition J. He is also member of the editorial advisory board of the International Journal of Image and Graphics and the Intelligent Automation and Soft Computing journal. He also served as Associate Editor of Simulation, the Journal of The Society for Computer Simulation.

Based on his work at the NCR, he received the NCR Inventor Award. He is also a recipient of the Systems Research Foundation Award for outstanding presentation in 1985 and the ISRAM best paper award in 1992. In 1994 he has been awarded the IEEE Computer Society Press outstanding referee award. He was also a coauthor of the best paper in the 2000 IEEE Canadian Conference on electrical and Computer Engineering. Dr. Kamel is recipient of the University of Waterloo outstanding performance award twice, the faculty of engineering distinguished performance award. Dr. Kamel is member of ACM, PEO, Fellow of IEEE, Fellow of the Engineering Institute of Canada (EIC), Fellow of the Canadian Academy of Engineering (CAE) and selected to be a Fellow of the International Association of Pattern Recognition (IAPR) in 2008. He served as consultant for General Motors, NCR, IBM, Northern Telecom and Spar Aerospace. He is co-founder of Virtek Vision Inc. of Waterloo and chair of its Technology Advisory Group. He served as member of the board from 1992 to 2008 and VP research and development from 1987 to 1992.

Fakhreddine Karray (*S*'89,M90,SM'01) received Ing. Dipl. in Electrical Engineering from University of Tunis, Tunisia (84) and Ph.D. degree from the University of Illinois, Urbana-Champaign, USA (89). He is Professor of Electrical and Computer Engineering at the University of Waterloo and the Associate Director of the Pattern Analysis and Machine Intelligence Lab. Dr. Karray's current research interests are in the areas of autonomous systems and intelligent man–machine interfacing design. He has authored more than 200 articles in journals and conference proceedings. He is the co-author of 13 patents and the co-author of a recent textbook on soft computing: Soft Computing and Intelligent Systems Design, Addison Wesley Publishing, 2004. He serves as the associate editor of the IEEE Transactions on Mechatronics, the IEEE Transactions on Systems Man and Cybernetics (B), the International Journal of Robotics and Automation and the Journal of Control and Intelligent Systems. He is the Associate Editor of the IEEE Control Systems Society's Conference Proceedings. He has served as Chair (or) co-Chair of more than eight International conferences. He is the General Co-Chair of the IEEE Conference on Logistics and Automation, China, 2008. Dr. Karray is the KW Chapter Chair of the IEEE Control Systems Society and the IEEE Computational Intelligence Society. He is co-founder of Intelligent Mechatronics Systems Inc. and of Voice Enabling Systems Technology Inc.