Artif Intell Rev
DOI 10.1007/s10462-012-9368-5

Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011

Christos-Nikolaos Anagnostopoulos · Theodoros Iliou · Ioannis Giannoukos

Abstract Speaker emotion recognition is achieved through processing methods that include isolation of the speech signal and extraction of selected features for the final classification. In terms of acoustics, speech processing techniques offer extremely valuable paralinguistic information derived mainly from prosodic and spectral features. In some cases, the process is assisted by speech recognition systems, which contribute to the classification using linguistic information. Both frameworks deal with a very challenging problem, as emotional states do not have clear-cut boundaries and often differ from person to person. In this article, research papers that investigate emotion recognition from audio channels are surveyed and classified, based mostly on extracted and selected features and their classification methodology. Important topics from different classification techniques, such as databases available for experimentation, appropriate feature extraction and selection methods, classifiers and performance issues are discussed, with emphasis on research published in the last decade. This survey also provides a discussion on open trends, along with directions for future research on this topic.

Keywords Speech features · Emotion recognition · Classifiers

C.-N. Anagnostopoulos (✉) · T. Iliou · I. Giannoukos
Cultural Technology and Communication Department, University of the Aegean, Lesvos isl., 81100 Lesbos, Greece
e-mail: canag@ct.aegean.gr

T. Iliou
e-mail: th.iliou@ct.aegean.gr

I. Giannoukos
e-mail: igiann@ct.aegean.gr

Published online: 09 November 2012

1 Introduction

1.1 Preface, scope and motivations of this survey

Communication is a fundamental faculty, based not only on linguistic statements but also on the emotional part. In the field of human-computer interaction (HCI), emotion recognition by the computer is still a challenging issue, especially when recognition is based solely on voice, which is the basic mean of human communication. Speech carries linguistic (explicit) information that is associated with emotions, along with paralinguistic (implicit) information, which can be extracted by speech processing methods. Linguistic information identifies qualitative patterns that the speaker has articulated, while paralinguistic information is usually measured by quantitative features describing variations in the way that the linguistic patterns (i.e words or phrases) are pronounced. The latter includes variations in pitch and intensity without linguistic information as well as voice quality and is related to spectral properties that cannot be correlated to word identity.

Numerous hi-level or low level acoustic features that are based mostly on elements such as pitch, energy, timing and intensity are usually proposed when trying to isolate emotion-specific information in the speech signal. In certain cases, those features are then reduced by feature selection methods and the final set is considered the input for the classification method that follows.

From the beginning of our survey, we must acknowledge that emotion recognition from speech is a rather difficult problem to solve, as sometimes even a human cannot easily classify natural emotions based on speech hue. It is therefore excessive to expect that machines can offer substantially correct classification. The purpose of this paper is to present a comprehensive survey of emotion recognition systems from speech in order to provide pattern recognition and speech processing researchers with basic information, theoretical background, materials and methods and current trends of this field. . In this survey three important issues of speech emotion recognition are presented: (1) available emotion databases and their usability in speech emotion recognition (2) various feature selection methods on previously extracted sound features and their specific contribution in speech emotion recognition performance and (3) numerous classifiers that have been used in speech emotion recognition portraying their classification rate as reported in the literature.

It should be noted that the purpose of this survey is to cover developments during the last decade in the field of emotion recognition from speech and not to compare the papers and declare “a winner” among them, as there is lack of uniformity in the way the methods are evaluated and assessed. Moreover, this overview does not provide an exhaustive listing of all papers, since there are papers and reports presenting overlapping information or slight modifications. The research literature related to building emotion recognition applications has been surveyed in other works. The aforementioned surveys focus on different aspects-as is expected in an interdisciplinary field-that range from psychology and neuroscience to computer science, engineering and education. A significant contribution of this survey that differentiates it from other survey papers is that it assembles the basic pieces of a rather complex pattern recognition problem in terms of computer science, including the non-linguistic vocalization recognition issue, which is not widely discussed among the researchers of this topic.

1.2 Definition of and models of emotions

The issue of defining emotions, distinguishing them from other affective states or traits and measuring them in a meaningful way has been a constant challenge for emotion researchers

in different disciplines of the social and behavioral sciences over a long period of time. Fontaine et al. (2010) proposed the following "working definition of emotion" for which there is increasing consensus in the literature. Emotions are episodes of coordinated changes in several components (including at least neurophysiologic activation, motor expression and subjective feeling but possibly also action tendencies and cognitive processes) in response to external or internal events of major significance to the organism. Adopting this theoretical approach, four dimensions are needed to satisfactorily represent similarities and differences in the meaning of emotions. In order of importance, these four dimensions are evaluation-pleasantness, potency-valence, activation-arousal, and unpredictability. From this 4-dimensional space, the research community of speech and video processing focuses mainly on 2-D representations forming the general idea of the "Emotion Wheel" (Cowie et al. 2000; Russell et al. 1989; Hanjalic and Xu 2005).

Several emotion wheels have been proposed based on pure appraisal dimensions, such as valence and arousal (Cowie et al. 2000; Russell et al. 1989). The Geneva Emotion Wheel (GEW) was the first one to present the dimensional layout of the emotion qualities on pure appraisal dimensions (control and valence) as well as the intensity of the associated subjective feeling (distance from origin). Apart from 2-D spaces, the study in Hanjalic (2006) experimented in 3-D space, where arousal, valence and control are used together. It should also be noted that although the OCC model (Ortony et al. 1988) is quite well-known in cognitive sciences, it is not widely adopted in computer science applications due to its great complexity.

Aiming to present a comprehensive and critical survey of up-to-date speech emotion recognition methods, the surveyed papers are discussed in two basic sections. In Sect. 2 the papers are categorized in sub-sections, according to their major methodology that includes issues such as linguistic and non-linguistic features, feature selection and classification. Critic assessments for each subsection, current trends and anticipated research in the above topics are discussed in Sect. 3.

2 Emotion recognition from speech

Speech emotion recognition is basically performed through pure sound processing without linguistic information. In terms of acoustics, speech processing techniques offer extremely valuable information derived mainly from prosodic and spectral features. Sometimes the process is assisted by Automatic Speech Recognition (ASR) systems, which contribute to classification using linguistic information. However, the use of ASR is limited due to fact that most of the experiments in the field have been assessed using databases of non-spontaneous and predefined speech and thus, there is no need for speech recognition. After sound processing and feature acquisition, it is quite common to follow a feature selection in search for the "golden set" of sound features. Finally, such a plethora of classification algorithms has been evaluated for speech emotion recognition, that attempting their comparison in this paper is, unfortunately, an impractical task. This is also due to the fact that there is a lack of uniformity in the way these methods are evaluated (different test sets, feature vectors and evaluation frameworks) and, therefore, it is inappropriate to make direct comparisons or explicitly declare which methods demonstrate the highest performance. In the next sections, a brief classification of papers that follow the basic processing pipeline (as highlighted in Fig. 1) are surveyed and categorized according to their major methodology for feature processing (with or without linguistic information), as well as their classification schema for emotion recognition.

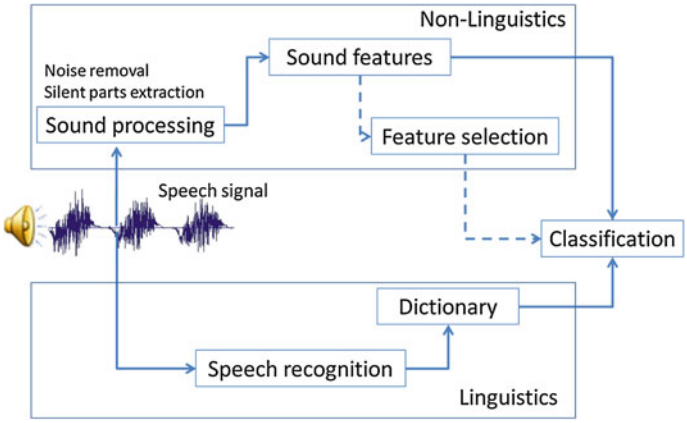

Fig. 1 Speech emotion recognition processing pipeline

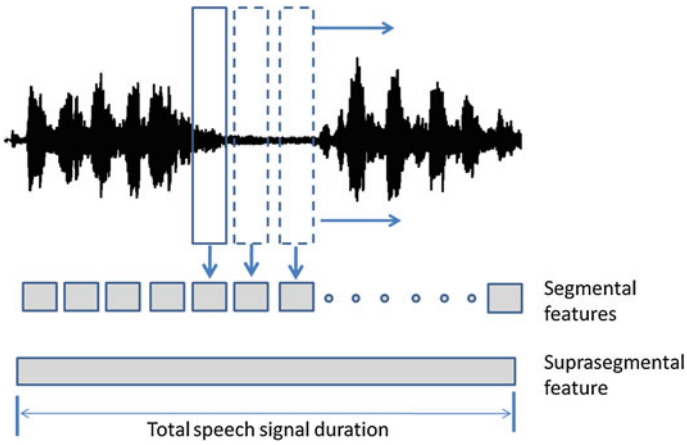

Fig. 2 Segmental and suprasegmental features in a speech signal

2.1 Creation of the feature vector

2.1.1 Non-linguistics features

Vector features are categorized as short-time (segmental) or long-time (suprasegmental) according to their temporal structure. Segmental features are calculated once for every small time frame (usually 25–50 msec using windowing techniques), allowing the analysis of their temporal evolution. In contrast, suprasegmental features are calculated over the entire utterance duration (as seen in Fig. 2). A quantitative feature-type-wise comparison between short-time and suprasegmental analysis is carried out for the recognition of interest in human conversational speech in Schuller and Rigoll (2009).

Moreover, vector features are also classified in two other distinctive classes, namely Low-Level-Descriptors (LLDs) and functionals. LLDs contain prosodic features, which are supra-segmental and spectral features and their derivatives that are segmental. The second class

Table 1 Speech features and their description

Features	Description
Mel-frequency cepstral coefficients (MFCCs), Linear prediction cepstral coefficients (LPCCs)	Derive from cepstrum, which is the inverse spectral transform of the logarithm of the spectrum Bogert et al. (1963)
Formants (spectral maxima or spectral peaks of the sound spectrum of the voice), log-filter-power-coefficients (LFPCs)	Derive from Spectrum
Noise-to-harmonic ratio, jitter, shimmer, amplitude quotient, spectral tilt, spectral balance	Are measurements of Signal (voice) quality
Energy, short energy	Are measurements of intensity
Fundamental frequency (pitch)	Are measurements of frequency
Temporal features (duration, time stamps)	Are measurements of time

Table 2 Speech feature categorization according to their temporal structure (suprasegmental vs. segmental) and parameterization (LLDs vs. functionals)

Low-level descriptors (LLDs)	Functionals (applied to LLDs)
Suprasegmental features	
Fundamental frequency (Pitch), energy, intensity, harmonics-to-noise ratio (HNR), shimmer, jitter, speech rate, normalized amplitude quotient, spectral tilt, spectral balance	Extreme values (maximum, minimum), means (arithmetic, quadratic, geometric), moments (standard deviation, variance, kurtosis, skewness), percentiles and percentile ranges, quartiles, centroids, offset, slope, mean squared error, sample values, time/durations
Segmental features	
Mel frequency cepstral coefficients (MFCCs), formant amplitude, formant bandwidth, formant frequency, log-filter power coefficients (LFPCs), linear prediction cepstral coefficients (LPCCs), line spectral pairs, short (Frame) energy, frame intensity	

(functionals) includes statistical features that derive from LLDs and therefore, they are suprasegmental features.

Specifically, LLDs include prosodic and spectral features, such as fundamental frequency (or pitch), energy, formants, Mel Frequency Cepstral Coefficients (MFCCs), Linear Prediction Cepstral Coefficients (LPCC), speaking rate, shimmer, jitter, voice quality parameters, while functionals include their statistics such as mean, maximum, minimum, change rate, kurtosis, skewness, zero-crossing rate, variance and so on. A description of speech features is given in Table 1, while in Table 2 those features are classified according to their temporal structure and parameterization.

The focus of the early studies was placed mainly on prosodic features in particular pitch, duration and intensity, while comparably small feature sets (10–100) were utilized (Nwe et al. 2003; Schuller et al. 2003; Lee et al. 2004; Wang et al. 2008). In more recent research, however, voice quality LLD features such as HNR, jitter, or shimmer, and spectral and cepstral measurements (e.g. MFCCs) have been extensively used (Schuller et al. 2007, 2009a; Lugger and Yang 2007a,b; Firoz Shah et al. 2009; Neiberg et al. 2006; Vlasenko et al. 2007;

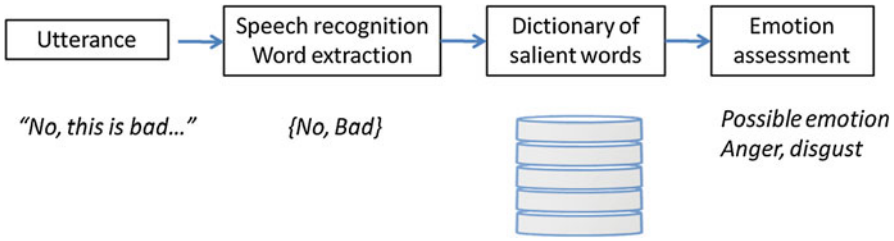

Fig. 3 A typical linguistic-information processing flow

Wenjing et al. 2009; Mishra and Sekhar 2009; Kostoulas et al. 2010; Anagnostopoulos and Iliou 2010; Wu and Liang 2011). Rhythm and sentence duration were included, along with classical measurements, such as pitch, energy and formants, as a classification feature in Yang et al. (2009a,b) and Jin et al. (2009). In addition, Non-Uniform Perceptual Linear Predictive features (UNPLP) (Zhou et al. 2009b), as well as Linear Predictive Cepstral Coefficients (LPCCs) (Schuller et al. 2003; Pao et al. 2007a; Fu et al. 2008a; Mao et al. 2009) accompany MFCCs in the feature set.

Functionals are also used to derive LLD statistics per utterance. Zero crossing rate was evaluated in Giannakopoulos et al. (2009), Atassi and Esposito (2008), Kostoulas et al. (2007), while extensive use of functionals is more recently reported in Schuller et al. (2010). Moreover, instead of feature selection and reduction of the feature space by well known methods (Schuller et al. 2006) propose the expansion of the set by generating novel features based on the existing ones. The feature set begins with dynamic Low-Level Descriptors such as intonation, intensity, formants and spectral information. Next, systematic derivation of prosodic, articulatory and voice quality high level operation is performed by descriptive statistical analysis. From that point, the feature set is enhanced with automatic feature alterations, in order to find an optimal representation within feature space in view of a target classifier. The search is performed following the principles of evolutionary programming. This completely different approach reported improvement of the classification performance compared to the authors' former works, which used two public databases.

Researchers tend to favor suprasegmental features (e.g prosodical or functionals) for the input feature vector, as these identify emotions better than segmental features (Sidorova 2007; Schuller et al. 2009b). To this end, traditional segmental features like MFCCs or LPCCs are transformed to suprasegmental parameterizations through long-time statistic processing, in order to be concatenated to the prosodic feature vector (Kwon et al. 2003; Vogt and André 2006; Schuller et al. 2005a). Free scientific software programs for speech processing, labeling, spectrographic analysis and pitch analysis in phonetics are available to the researchers. Among the most popular toolkits, PRAAT (http://www.fon.hum.uva.nl/praat) and SNACK (http://www.speech.kth.se/snack/) are designed to be used with scripting languages that enable the creation of multi-platform audio processing.

2.1.2 Linguistic information

As already mentioned, linguistic information schemas consist of ASR systems that identify specific words or phrases which can be correlated to an emotional state. It is clear that, beside the need for effective speech recognition, the existence of an updated dictionary is of paramount importance for the successful implementation of linguistics in emotion recognition. A typical processing flow is shown in Fig. 3.

In an early research Batliner et al. (2003) suggested that apart from prosodic features in the speech signal, syntactic and behavioral hints like repetitions in a dialogue and part-of-speech features, should be taken into consideration. According to this article, repetitions and reformulations (the use of different words to convey the same content) have been found to be indicators of changing speaker attitude, such as increasing frustration and anger. Moreover, the part-of-speech (POS) of each word representing a rough syntactic structure was categorized in 6 classes: noun, inflected adjective, non-inflected adjective, copula, verb and pronoun. The POS of each word has been annotated manually in a lexicon that contains all word forms found in the database. The research revealed that nouns and adjectives are more useful in emotion categorization arguing the point that generally, content words are more salient and more prone to be emotionally marked than function words (i.e. verbs).

In Lee and Narayanan (2005), a combination of three sources (acoustic, lexical and discourse) was used for emotion recognition. To capture emotion information at language level, an information-theoretic notion of emotional salience in two hyper-classes (negative and non-negative emotion) was introduced. The salience of a word in emotion recognition can be defined as mutual information between a specific word and emotion category. For instance, the word "wrong" would generate a salience output towards the negative emotion class, while the word "exactly" would indicate a non-negative emotion. A list of salient words was collected and if the words in a given utterance match those salient words, an output of "1" from those words is generated according to the two emotion classes. The results show that combining all the information, rather than using only acoustic information, improves emotion classification by 40.7 % for males and 36.4 % for females.

Similarly, Litman and Forbes-Riley (2004) examine the utility of speech and lexical features for predicting student emotions in computer-human tutoring dialogues. Emotion annotation was performed for negative, neutral, positive and mixed emotions. Prosodic features are then extracted from the speech signal and lexical items (words) from recognized speech. The results yield a 19–36 % relative improvement in error reduction over a baseline of previous experiments. In Forbes-Riley and Litman (2004) the authors also investigated the role of context information (e.g., subject, gender and turn-level features representing local and global aspects of the dialogue) on audio affective recognition. They discuss a scheme for manually annotating student turns in a human-human tutoring dialogue corpus for 3 intuitively perceived emotion scales (negative, neutral and positive). Thus, besides lexical items, a hint of the emotion can come from other knowledge sources, such as the gender of the student, a long pause before answering a question or a sudden change in speech loudness and tempo.

Voice stress analysis procedures attempt to use low-level indicators of stress as indirect indicators of deception. Based on the above, Graciarena et al. (2006) propose a combined approach using linguistic information along with prosodic measurements to detect deception. Additionally, in a recent work Schuller et al. (2009c) improve the automatic recognition of emotion from spoken words by applying vector space modeling versus string kernels. Apart from the spoken content, the authors integrated Part-of-Speech and higher semantic tagging in their analysis.

Ijima et al. (2009) presented a technique which can obtain linguistic information using keywords. However, all keyword-based systems demonstrate several problems, such as degree of ambiguity in emotional keywords and lack of affect-related semantic and syntactic knowledge base. In this context, Semantic Labels (SLs) have been introduced from Wu et al. (2006). Moreover, in a very recent paper Wu and Liang (2011) fuse the results of semantic label classification with acoustic-prosodic information to boost emotion recognition in affective speech. Semantic labels were derived from the Chinese knowledge base HowNet

(http://www.keenage.com/) and used to automatically extract Emotion Association Rules (EARs) from the recognized word sequence of affective speech.

2.1.3 Non-Linguistic Vocalizations

There is also a special sub-category in non-linguistic information that relates to human vocalizations (often referred to as non-linguistic vocalizations). Laughs, cries, sighs, yawns and other similar vocal outbursts seem at first to be good examples of expressions of discrete (although not necessarily basic) emotions. A funny joke elicits amusement, which produces a laugh; a loss elicits sadness, which produces crying; an uninspired lecture elicits boredom, which produces a yawn. However, it is uncertain whether all vocalizations are linked to specific, discrete states (Russell et al. 2003).

As a result, few efforts have been reported towards the automatic recognition of non-linguistic vocalizations such as laugher (Petridis and Pantic 2008), cries (Pal et al. 2006), emotional bursts (Schroder 2003) and coughs (Matos et al. 2006). In addition, certain vocal expressions and prosodic cues have been associated with specific mental illnesses such as Obsessive Compulsive Disorder (Aigner et al. 2007), Tourette Syndrome (Calder et al. 2001) or depression (France et al. 2000). These expressions usually manifest themselves as vocal cues of fear and sadness, usually accompanied by corresponding facial expressions. These vocal cues (or tics) comprise of compulsive barking and grunting noises, frequent throat clearing, coughing or sniffling, echolalia (vocal tics characterized by repeating words), and/or coprolalia (vocal tics characterized by repeating or shouting obscene words). However, no effort towards automatic effective emotion recognition analysis based on vocal outbursts has been reported so far.

2.1.4 Feature selection

Prior to classification, feature selection, also known as variable selection or feature reduction, is often used in speech emotion recognition in order to speed up the learning process and minimize the problem known as "the curse of dimensionality". Popular feature selection methods that have been implemented include Principal Component Analysis (PCA) (Zhou et al. 2009b; Schuller et al. 2005b; Wang et al. 2010; Wagner et al. 2005) and Canonical Correlation Analysis (CCA) (Cheng et al. 2009). Moreover, Correlation-based Sub Set Evaluators have also been used for feature selection, where several search methods evaluate a subset of features for the optimal subset. Such search methods include BestFirst (Kostoulas et al. 2010; Anagnostopoulos and Vovoli 2010), correlation-based analysis (Vlasenko et al. 2007; Vogt and André 2009), Genetic Algorithms (Wang et al. 2008), Support Vector Machine-Sequential forward floating search (Schuller et al. 2005a,c), Mutual Information (MI) between the class Y and an attribute X (Schuller et al. 2005b; Hoch et al. 2005) and the Sequential Floating Forward Selection (SFFS) algorithm (Atassi and Esposito 2008). Many of the aforementioned works have been carried out using freely available, open-source software platforms like WEKA (http://www.cs.waikato.ac.nz/ml/weka/) or RapidMiner (http://rapid-i.com/content/view/181/196/).

2.2 Classification

2.2.1 Classification frameworks

Usually, classification evaluations are carried out using a single database or dataset. In this case, several testing frameworks appear based on the dependency or not on the speaker

(speaker dependent/independent) as well as the context (context dependent or independent). The "independent" frameworks provide more reliable evaluation and thus, are more commonly used in the literature, sometimes with different names, such as leave-one-speaker-out (LOSO) or leave-one-speaker-group-out (LOSGO). However, every speech database is created on the basis of fixed recording conditions and noise levels and specific room acoustics, while the data is recorded only in one language. Indicatively, in Yang et al. (2009a) it is demonstrated that the features selected to build a classifier are not that robust for speech emotion recognition in a different language. Moreover, another simplification that characterizes most of the classification frameworks is that systems are usually trained and tested using the same database.

In order to avoid the simplification of training the system in the same database, Schuller et al. (2010) propose cross-corpora evaluation to increase independence between training and testing sets. Specifically, they demonstrate results employing six standard databases (AVIC, DES, EMO-DB, eNTERFACE, SmartKom, SUSAS) in a cross-corpora and multilingual evaluation experiment. The research discovered similarities among databases, a fact that indicates what kind of databases can be combined to obtain further training material for emotion recognition systems and thus reduce the problem of data sparseness.

2.2.2 Single classifiers used for speech emotion recognition

For emotional state modeling, a variety of pattern recognition methods are utilized to construct a classifier, such as Gaussian Mixture Models (GMMs), Hidden Markov Models (HMMs), Support Vector Machines (SVMs), Artificial Neural Networks (ANNs), Decision Trees or k-Nearest Neighbor distance classifiers (kNNs).

Among all classifiers, Gaussian Mixture Models (GMMs) (Lugger and Yang 2007a; Neiberg et al. 2006; Zhou et al. 2009a; Kostoulas et al. 2010; Zhou et al. 2009b; Pao et al. 2007a; Graciarena et al. 2006) have been studied the most. GMMs are probabilistic models for density estimation using a convex combination of multi-variate normal densities. They are very efficient in modeling multi-modal distributions (Douglas-Cowie et al. 2007) and their training and testing requirements are considerably fewer than the requirements of a general continuous HMM. Therefore, GMMs are more appropriate for speech emotion recognition for global feature extraction, as prosodic features are usually processed on a frame-level basis (suprasegmental features). Similar to many other classifiers, the definition of the optimum number of Gaussian components is a difficult task that cannot be addressed uniformly in the literature. The applicability of Variational Gaussian Mixture Models (VGMMs) to emotion speech recognition was also demonstrated in Mishra and Sekhar (2009).

Hidden Markov Models (HMMs) are considered to be a common classification/modeling technique for speech emotion recognition (Schuller et al. 2003; Pao et al. 2007a; Ijima et al. 2009; Yun and Yoo 2009; Yu 2008; Fu et al. 2008b; Nogueiras et al. 2001). Usually, each emotion is modeled by a single state Hidden Markov Model (HMM) that is trained by maximizing the minimum separation margin between emotions, while the margin is scaled by a loss function. In contrast to GMMs, HMMs are stochastic processes which consist of a first-order Markov chain whose states are hidden from the observer. Since the association with each state is a random process that generates the observation sequence, the hidden states of the model capture the temporal structure (segmental features) of the data. A critical design issue of an HMM classifier is the determination of the optimal number of states, the type of the observations and the optimal number of observation symbols (for discrete HMM) or the optimum number of Gaussian components (for continuous HMM). As in the above studies,

HMMs have been used in stress state recognition (Nwe et al. 2003; Fernandez and Picard 2003).

In contrast, Support Vector Machines (SVMs) have been used more recently and seem to be promising as a classification schema for emotion recognition in speech, as assessed in many papers (Schuller et al. 2005a,c, 2009c, 2010; Wang et al. 2008; Graciarena et al. 2006; You et al. 2006; Vlasenko et al. 2007; Luengo et al. 2010; Wu et al. 2009). SVMs offer specific advantages over GMM and HMM including the global optimality of the training algorithm and the existence of excellent data-dependent generalization bounds (Mishra and Sekhar 2009). On the other hand, their success in non-separable cases is relatively heuristic. There is no systematic way of choosing the kernel functions and as a result, the separation of the transformed features is not always guaranteed. In fact, in the problem of emotion recognition from speech, perfect separation of the training data is not correct in order to avoid over-fitting. Moreover, Yang et al. (2009a) studied the problem using Twins Support Vector Machines (TWINsSVM). Comparisons on classification algorithms between TWINsSVM and standard SVMs revealed that TWINsSVM can achieve marginally higher performance.

Several Computational Intelligence classifiers were also reported in the literature, such as Artificial Neural Networks (ANNs), Fuzzy Sets (Yang et al. 2009b) and Evolutionary Algorithms, with the latter being a good feature selection method (Wang et al. 2008). The list of ANNs includes Multi Layer Perceptrons (MLPs) (Anagnostopoulos and Vovoli 2010; Firoz Shah et al. 2009; Fu et al. 2008a), Probabilistic Neural Networks (Cen et al. 2008), Vector Quantization networks (Wenjing et al. 2009) and Deep Neural Networks (Stuhlsatz et al. 2011). In addition, MLP various architectures have been tested, like All-Class-in-One-Network (ACON), where all the classes are placed in a single network and One-Class-in-One-Network (OCON), where an individual single network is responsible for each and every class (Wang et al. 2010). Specific advantages of ANNs include increased effectiveness in modeling nonlinear mappings and better classification performance than HMM and GMM when the number of training examples is relatively small. However, there is no common rule for setting the optimal ANN topology, which is usually defined ad-hoc. The topology, along with the selection of the activation functions, the number of training epochs, the learning rate and the validation methods, affects the reported results in a way that makes performance comparisons an extremely hard task.

Decision Trees have also been assessed as classifiers with the well-known C4.5 algorithm leading the relevant studies (Rong et al. 2007; Kostoulas and Fakotakis 2006), while the Random Forest (RF) classifier was assessed in Rong et al. (2007), Iliou and Anagnostopoulos (2009). Random forest (or random forests) is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the responses provided by individual trees. Among the advantages of RFs that can be applied in emotion recognition from speech is the fact that they run effectively on large databases, handling thousands of input features without feature detection. However, few researchers choose to implement RF, because the resulting classifications are difficult for humans to interpret as RFs present a tendency for data over-fitting.

As far as the linguistic content classification is concerned, extensive research was done using the Bag of Words (BoW) and N-Grams model. BoW is a form of vector space modeling and a well-known numerical representation form of text in automatic document categorization, information retrieval and natural language processing. Specifically, experiments on the FAU Aibo Emotion Corpus have shown surprisingly low performance degradation with real ASR over transcription based emotion recognition (Schuller et al. 2009c). In the experimental results, Bag of Words (BoW) dominated all other modeling forms based on spoken content. Each word in the vocabulary adds a dimension to a linguistic vector representing the term

frequency within the actual utterance. This method was successfully ported to recognize sentiments or emotion in Schuller et al. (2005a, 2009a). On the other hand, N-Grams model has been frequently employed in statistical natural language processing (Manning and Schütze 1999). An n-gram model is a type of probabilistic model for predicting the next item in a sequence of n items. In the cases examined, items were phonemes, syllables, letters or words (Ang et al. 2002; Lee et al. 2002; Devillers et al. 2003).

Table 3 highlights the best performance of single classification schemas as reported in the papers surveyed above. It should be noted that these rates are given for overview purposes and not for direct comparison, since there is a lack of uniformity in the way methods are evaluated.

2.2.3 Hybrid classifiers, ensembles, voting schemes

Several classifiers may not perform well on all emotional states. For example, a GMM-based classifier may fail to correctly recognize the neutral emotion, while the MLP-based classifier is clearly superior in neutral emotion recognition. Therefore, hybrid classifiers and ensembles were proposed in order to achieve higher recognition performance than individual classifiers. Table 4 presents an overview of performance issues, while a typical architecture of such schemas is shown in Fig. 4.

In Pao et al. (2007b) a multiple classifier system is established to achieve the best possible classification performance for emotion classification in Mandarin speech emotional corpus, which includes five emotions: anger, happiness, sadness, boredom and neutral. The classifiers that were investigated by the authors include K-Nearest Neighbor (KNN), Weighted KNN (WKNN), Weighted Average Patterns of Categorical KNN (WCAP), Weighted Discrete KNN (W-DKNN) and Support Vector Machine (SVM). Three classifier combining rules were tested, namely majority voting, minimum misclassification and maximum accuracy methods and the combined results outperformed the accuracy of single classifiers.

Voting schemes are also reported for speech emotion recognition fusing utterance classification and frame-based classification. Specifically, in the research of Shami and Kamel (2005), each utterance is viewed as a series of distinctive voiced parts and not as a single entity. The voiced parts undergo segmentation, followed by statistical measurements of spectral shape, intensity and pitch contours. Final utterance classification is performed by combining the segment and utterance classification decisions using a fixed voting scheme. Moreover, other selections of ensemble techniques have been applied, such as Boosting, Bagging, Multiboosting, and Stacking (Schuller et al. 2005a,c; Rong et al. 2007; Morrison et al. 2007).

Linguistic features were also combined in hybrid classification systems. Such results are reported in the recent paper of Wu and Liang (2011). In this work, GMM, SVM, and MLP are employed in a meta-decision tree (MDT) architecture to model the acoustic-prosodic information based on speech features. Speaker-independent experimental results revealed that the emotion recognition performance based on the MDT model improved by 3.5 % (overall 83.5 %), when linguistic features were used. Similarly in Schuller et al. (2004), linear classifiers, Gaussian Mixture Models, Neural Nets and Support Vector Machines were combined for acoustic feature classification, while a Belief Network spotted emotional key-phrases from an automatic speech recognition (ASR) engine based on Hidden Markov Models. Finally, the two information sources were integrated in a soft decision fusion using a Neural Net to improve the overall performance up to 8.0 %.

Table 3 A brief overview of classification performance in single classifiers

Classifier	Performance	Reference
SVMs	87.5 % in Berlin EMO database	Schuller et al. (2005a)
	70.3 %, 7 emotions (large unknown dataset)	Schuller et al. (2005c)
	up to 81 % in several cross-corpus experiments with varying number of classes	Schuller et al. (2010)
	up to 88.15 %, 6 emotions (unknown dataset)	Wang et al. (2008)
	∼63 % in a corpus that includes deceptive and non-deceptive speech	Graciarena et al. (2006)
	∼90 % in Berlin EMO database and 83 % in SUSAS	Vlasenko et al. (2007)
	78 % in Berlin EMO database	Luengo et al. (2010)
	88.6 % in Berlin EMO database	Wu et al. (2009)
	89 % in Berlin EMO and DSPLAB databases	Yang et al. (2009a)
	75.33 % (speaker independent), 4 emotions in 2 unknown datasets	Wu and Liang (2011)
	78.16 % (speaker independent with linguistic information), 4emotions in 2 unknown datasets	Wu and Liang (2011)
	71.85 %, 6 emotions (unknown dataset)	Morrison et al. (2007)
GMMs	∼62 % in a corpus that includes deceptive andnon-deceptive speech	Graciarena et al. (2006)
	86 % in Chinese-LDC	Zhou et al. (2009b)
	81 % in Berlin EMO database (speaker independent)	Atassi and Esposito (2008)
	74.6 % in Berlin EMO database (speaker independent)	Lugger and Yang (2007a)
	70.3 %, 5 emotions (unknown dataset)	Pao et al. (2007a)
	90 % in 3 classes (neutral, emphatic, negative)	Neiberg et al. (2006)
	∼50 % in FAU Aibo Emotion Corpus	Kostoulas et al. (2010)
	∼63 % in Berlin EMO database	Mishra and Sekhar (2009)
	68.73 % (speaker independent), 4 emotions in 2 unknown datasets	Wu and Liang (2011)
	72.61 % (speaker independent with linguistic information), 4 emotions in 2 unknown datasets	Wu and Liang (2011)
HMMs	89 % in Berlin EMO database	Yun and Yoo (2009)
	∼87 %, 5 emotions (in two unknown datasets)	Yu (2008)
	∼81 %, 5 emotions (unknown dataset)	Ijima et al. (2009)
	62.5 %, 5 emotions (unknown dataset)	Pao et al. (2007a)
	78.4 % in Berlin EMO database (speaker independent)	Fu et al. (2008b)
	∼80 % in ELSA multi-lingual emotional speech database	Nogueiras et al. (2001)
	86 %, 7 emotions (unknown dataset)	Schuller et al. (2003)
ANN	68.5 %, 4 emotions (unknown dataset)	Firoz Shah et al. (2009)
	63.3 % in Berlin EMO database and 61.4 % (unknown dataset)	Fu et al. (2008a)
	∼60 % (speaker dependent) and 55 % (gender dependent) in LDC emotional prosody speech-transcripts database	Cen et al. (2008)
	71.4 %, 4 emotions (unknown dataset)	Wenjing et al. (2009)
	47 % in Berlin EMO database (speaker dependent but utterance independent)	Anagnostopoulos and Vovoli (2010)
	52–62 %, 6 emotions (undefined dataset)	Wang et al. (2010)
	83.2 and 55 % (speaker dependent and independent) in Berlin EMO database	Iliou and Anagnostopoulos (2009)
	69.86 % (speaker independent), 4 emotions in 2 unknown datasets	Wu and Liang (2011)
	71.87 % (speaker independent with linguistic information), 4 emotions in 2 unknown datasets	Wu and Liang (2011)
	65.37 %, 6 emotions (unknown dataset)	Morrison et al. (2007)

Table 3 continued

Classifier	Performance	Reference
C4.5	76.5 %, 3 emotions in two unknown datasets	Rong et al. (2007)
	85.4 % (speaker dependent) in LDC emotional prosody speech-transcripts database	Kostoulas and Fakotakis (2006)
	61.5 % in Berlin EMO database	Schuller et al. (2005a)
	50 %, 7 emotions (large unknown dataset)	Schuller et al. (2005c)
RF	80.6 %, 3 emotions in two unknown datasets	Rong et al. (2007)
	77.2 and 48 % (speaker dependent and independent) in Berlin EMO database	Iliou and Anagnostopoulos (2009)
	67.36 %, 6 emotions (unknown dataset)	Morrison et al. (2007)
k-NN	72.2 %, 5 emotions (unknown dataset)	Pao et al. (2007a)
	61.83 %, 6 emotions (unknown dataset)	Morrison et al. (2007)

Table 4 A brief overview of performance in ensembles, voting and hybrid classifiers

Combination of	Overall performance	Reference
C4.5, RF	78.1 %, 3 emotions in two unknown datasets	Rong et al. (2007)
SVM, GMM, MLP	80 % (speaker independent), 4 emotions in 2 unknown datasets	Wu and Liang (2011)
SVM, GMM, MLP	83.55 % (speaker independent with linguistic information), 4 emotions in 2 unknown datasets	Wu and Liang (2011)
SVM, K-NN	83.8 %, 5 emotions (unknown dataset)	Pao et al. (2007b)
SVM, K-NN	87 % in KISMET database	Shami and Kamel (2005)
SVM, K-NN, Naïve Bayes, C4.5, ANN	80.5 % in Berlin EMO database	Schuller et al. (2005a)
SVM, K-NN, Naïve Bayes, Boosted C4.5	71.62 %, 7 emotions (large unknown dataset)	Schuller et al. (2005c)
SVM, MLP, K-NN, RF	73.3 %, 6 emotions (unknown dataset)	Morrison et al. (2007)

3 Discussion

3.1 Emotional datasets/databases

Surveying the literature, it becomes evident that emotion recognition in speech is mostly assessed using digital sources that are datasets rather than databases. Datasets are small-scale collections of material created to focus on a specific research and most importantly they are not widely available. Collections that are available to the community tend to fulfill requirements related to validity and generalization and, therefore, the term "database" is the most appropriate for them.

Generally, it is extremely difficult to produce a database representing the natural speech of a man or a woman in completely natural conversation. Many examples of humans talking exist, but very few of them illustrate speech in a natural environment. In the latter case, some databases use corpora (i.e. large collections) of spontaneous speech, usually consisting of clips from live television, radio programs or call centers, with natural speech recorded in

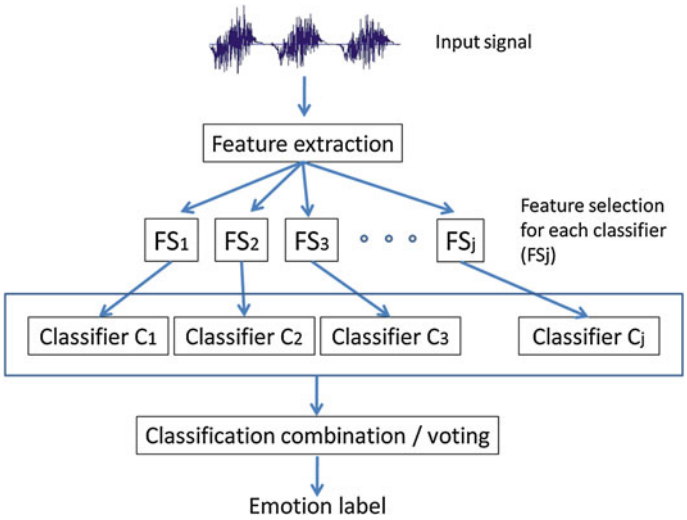

Fig. 4 A typical architecture for a combined classifier or a voting scheme

real-world situations. On the other hand, such databases are not distributed easily, since their assessment and processing could raise serious ethical or copyright issues.

Thus, in most cases, speech databases/datasets use acted speech, since the easiest way to collect emotional speech is to have actors simulate it. However, some questions are raised related to the naturalness of the outcome. There are many reasons to suspect that there are significant differences between acted and spontaneous speech. Actors often simply read the utterances or the passages, failing to wholeheartedly participate in their role. This could easily lead to the recoding of inaccurate characteristics in the speech signals. Moreover, actors may not capture the original context-related real-world emotions or exaggerate in their acting, making emotion recognition in acted speech easier than in spontaneous speech (Vogt and André 2005).

Table 5 summarizes the emotional speech databases or corpora that, to our knowledge, are available in the web (either with full free access or under license agreement) along with the respective access link. It should be noted that there are numerous smaller experimental datasets, which are not publicly available.

As mentioned above, capturing a faithful, detailed record of human emotion, as it appears in real life, is an incredibly challenging task. Practically most of the databases which have been provided by different sites do not contain realistic, non-prompted speech but prompted or acted speech. The assembly of databases (or datasets) has not traditionally been considered a high-profile or intellectually challenging area. Focus is explicitly placed in good quality recording and large samples that usually contain high arousal emotions (e.g. anger, sadness), while real human emotions are left relatively off-focus. Figure 5 shows that databases include intense emotions more frequently than low arousal ones.

Lately, we are happy to see a wider range of emotions covered, more elicited or even spontaneous sets of many speakers and larger amounts of instances (5–10 k) of more than 100 subjects in recent database developments. Another recent trend in the creation of databases is the recording of induced emotions from human subjects. In order to capture spontaneous reactions and emotion from human subjects that are unaware of being observed, specially designed scenarios are created to elicit an emotional state from the subject. This could be

Table 5 Databases of emotional speech

Name	Link
Socrates emotional speech database	http://www.wcl.ece.upatras.gr/ai/resources/demo-emotion-recognition-from-speech
DSPLAB emotional speech	http://wwwbox.uni-mb.si/eSpeech/speech.html
BELFAST naturalistic emotion database	http://www.idiap.ch/mmm/cor\discretionary-pora/emotion-corpus
Kids audio speech corpus	http://techexplorer.cusys.edu/show_NCSum.cfm?NCS=258629
LDC Emotional prosody speech-transcripts	http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002S28
Speech under simulated and actual stress (SUSAS)	http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC99S78
KISMIT database	http://www.ai.mit.edu/projects/sociable/expressive-speech.html
Bavarian archive for speech	http://www.bas.uni-muenchen.de/Bas/
Berlin database of emotional speech	http://pascal.kgw.tu-berlin.de/emodb/index-1280.html
ELRA-GRONINGEN database	http://catalog.elra.info/product_info.php?products_id=61
ELSA, Multi-lingual emotional speech database	http://universal.elra.info/product_info.php?cPath=37_39\&products_id=62
ELSA, emotional speech corpus	http://universal.elra.info/product_info.php?cPath=37_39\&products_id=115
ELSA, Danish emotional speech database	http://universal.elra.info/product_info.php?cPath=37_39\&products_id=78
RUSLANA database	http://www.accenture.com/NR/rdonlyres/A149A081-DB75-44C6-B250-F844E0C0C7C5/0/RuslanaUpdated.pdf
FAU Aibo emotion corpus	http://www5.informatik.uni-erlangen.de/en/our-team/steidl-stefan/fau-aibo-emotion-corpus/
VAM audio spontaneous speech database	http://emotion-research.net/download/vam
Simulated emotion speech corpus	http://www.springerlink.com/content/lg7655w4061ng6t1/
Interactive emotional dyadic motion capture	http://sail.usc.edu/iemocap/
HUMAINE database	http://universal.elra.info/product_info.php?cPath=25\&products_id=2063
SmartKom	http://www.smartkom.org
AVIC	Schuller and Rigoll (2009)
CASIA Mandarin emotional corpus	http://www.chineseldc.org
eNTERFACE	http://www.enterface.net/

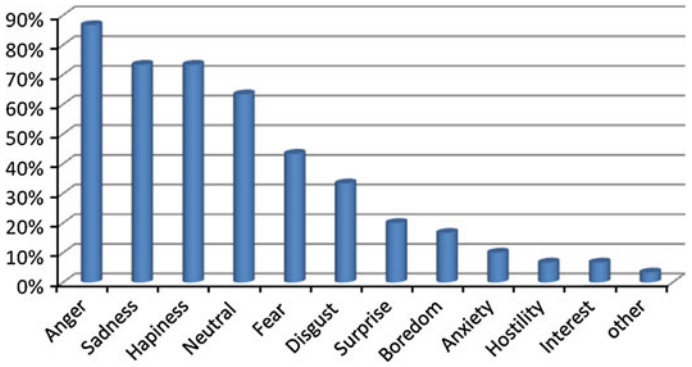

Fig. 5 Appearances of emotions in databases ("other" emotions include: Pride, elation, satisfaction, relief, hope)

achieved through an interaction with a specific content (or qualified actor) that gradually draws the human subject into a predefined emotional state. Of course, the effectiveness of this method at inducing the expected (or "target emotion") varies according to the emotional reaction of the participant in the experiment and their appraisal of the situation.

Very good reviews of emotional databases are given in Douglas-Cowie et al. (2003), Cowie et al. (2005) and Navas et al. (2006). In these reviews, it is mentioned that the payoff of a correctly created database would be tremendous and that technology is capable of keeping the research going in spite of all the difficulties. It is also noted that samples of correctly developed speech databases should be exchanged and cross-validated by the researchers in order to tackle the problem of multilingual context (Hozjan and Kacic 2006). The above issue is also mentioned in Schuller et al. (2010), along with directions for future research. According to the latter study, a substantial body of future research should highlight issues like multilingual context and cultural differences in expressing and perceiving emotion.

3.2 Speech features

The identification of emotion-related speech features is extremely challenging in audio analysis. In the real world, humans are able to detect spontaneous linguistic and paralinguistic information or even a combination of these, due to their remarkable ability to interpret emotional expressions. In the digital world though, the computer has to split the above tasks and deal with them using separate methodologies.

Conversely, for the non-linguistic channel of information, optimal sets of voice parameters to discriminate emotions are not yet identified. It seems that, according to the emerging methodologies, there is a strong interest in the scientific community to correlate some basic emotions with prosody features, such as pitch, MFCCs, formants and energy measurements. Although a number of paralinguistic features have been proposed in the literature surveyed above, the optimal feature set has not yet been established and the researchers strive to approximate the problem by segmenting the procedure in smaller problems rather than addressing the issue as a whole.

Considering the results reported in the papers that incorporate linguistic information, the main problem relates to the fact that automatic speech recognition technology has not yet reached the level of maturity required to perform consistently well in spontaneous speech. Despite the latest developments in the field, automatic extraction, separation and identification

of spoken words from spontaneous speech, especially if this speech is pregnant with emotions, is still a difficult problem. Moreover, the performance of Automatic Speech Recognition (ASR) systems declines if speech is not articulated according to specific rules. A study on adapting an ASR to emotional speech has been reported in Athanaselis et al. (2005). Linguistic information is considered successful under the assumption of perfect automatic speech recognition and sufficiently rich vocabulary in the lexicon. Exploitation of on-line knowledge sources without domain specific model training was recently proposed as an effective alternative in order to cope with out of-vocabulary events (Schuller et al. 2009d).

Also, assuming successful speaker recognition, an emotion recognition system from speech may be implemented as a combination of a speaker identification system followed by a speaker-dependent emotion recognition system. This is due to the fact, that speaker-dependent emotion classification is generally easier than speaker-independent classification.

Finally, the association between linguistic content and emotion is strongly language dependent, making the generalization from one language to another very tricky as reported in an excellent survey of affect recognition methods (Zeng et al. 2009). A very interesting research related to language dependence for emotion expression was initiated in Wierzbicka (2009). According to this study, bilingual people know well that when they try to describe the same experience in two different languages they are often forced to present it differently in each, because emotion words in the two languages may not match. For this reason, it is proposed that emphasis placed on how the use of a methodology developed in linguistic semantics known as NSM (Natural Semantic Metalanguage) can help us understand human emotions better. The goal of NSM is to make possible the study of human emotions from a genuinely cross-linguistic, cross-cultural and psychological perspective and, in this way, to open up new possibilities for the scientific understanding of subjectivity and psychological experience.

3.3 Classifiers

In the last decade, numerous studies that attempt to improve on features and propose effective classification schemas have been produced. However, until very recently, the researches usually made use of small, preselected, prototypical and often non-spontaneous emotional data sets and therefore comparability of results was inapplicable. An even bigger problem is the lack of exact reproducibility of the results, since results are reported on randomly partitioned sets of one emotional speech database in each study.

The fusion of the results of different classifiers in the decision level is also recognized as a problem. Each classifier for each channel of information is usually trained and optimized locally before the synthesis, making the results suboptimal. As identified by Lee and Narayanan (2005), further improvements can be made if the parameters of combined classifiers are globally optimized by jointly training the whole system. They add that the combination/fusion problem involving different information sources is an open question that continues to be tackled by the data fusion, signal processing, and machine learning communities. This problem is applied to classification fusion in paralinguistic methods as well.

There have been some attempts to provide researchers with a common benchmark and evaluation framework. A first cooperative experiment is found in the CEICES initiative (Batliner et al. 2006), where seven sites compared their classification results under the exact same conditions and pooled their features together for one combined unified selection process. This comparison was not fully open to the public, which motivated the INTERSPEECH 2009 Emotion Challenge (Schuller et al. 2009e) to be conducted with strict comparability, using the same database and common classification tools, such as Hidden Markov Model

Toolkit (http://htk.eng.cam.ac.uk/docs/docs.shtml) and WEKA (http://www.cs.waikato.ac.nz/ml/weka/).

The INTERSPEECH 2009 Emotion Challenge was the first open public evaluation of speech-based emotion recognition systems with strict comparability, where all participants were using the same corpus (FAU Aibo Emotion Corpus). Three sub-challenges (Open Performance, Classifier, and Feature) addressed the classification of five non-prototypical emotion classes (anger, emphatic, neutral, positive, remainder) or two emotion classes (negative, idle). The Open Performance Sub-Challenge allowed contributors to find their own features with their own classification algorithms abiding by the definition of test and training sets. In the Classifier Sub-Challenge, participants designed their own classifiers and had to use a large set of standard acoustic features, computed with the openSMILE toolkit (http://fr.sourceforge.jp/projects/sfnet_opensmile/). In the Feature Sub-Challenge, participants were encouraged to design 100 best features for emotion classification to be tested by the organizers in equivalent setting.

The organizers provided baselines (67.7 % unweighted average recall for the 2-class problem and 38.2 % unweighted average recall for the 5-class problem) under the condition of using WEKA and HTK toolkits, so that the results were reproducible. The metric of unweighted average recall (UAR) was selected, because it reflects better the imbalance among classes (usually there is a high percentage of neutral speech but sparse instances of diverse non-neutral examples). In the Open Performance Sub-Challenge, the best result in the two-class task reached 70.29 % UAR (Dumouche et al. 2009), while the best result in the five-class task was 41.65 % UAR (Kockmann et al. 2009). The Classifier Sub-Challenge winners were Lee et al., whose performance measure of the UAR percentage on the evaluation data set improved the baseline model by 3.3 % (Lee et al. 2009). In contrast, no award was given in the two classes Classifier Sub-Challenge and in the Feature Sub-Challenge, since the participants in these sub-challenges did not exceed the baseline results. A detailed analysis on the Emotion Challenge can be found in Schuller et al. (2011).

4 Conclusions

Research papers that investigate emotion recognition from audio channels were surveyed and classified mostly based on: (i) the features extracted and selected for training the classifiers (linguistic or non-linguistic) and (ii) their classification methodology. It should be emphasized that there is a lack of uniformity in the way methods are evaluated and, therefore, it is inappropriate to make direct comparisons and to explicitly declare which methods demonstrate the highest performance. Indeed, one of the main conclusions of this survey is to highlight that the evaluation of the proposed methods is often not performed in common test sets consequently a common reference point for algorithmic assessment cannot be achieved. Attempting to provide some solid conclusions that can be highlighted from this study, we can say that:

(1) Even people are sometimes confused when deciphering the emotional states of other individuals, so it is obvious that feature sets and classification methods which deal with this vague problem need to be further explored and studied.
(2) The “golden set” from an endless list of non-linguistic features has not been found yet. The well-defined and strictly pre-arranged testing environments of the Emotion Challenges as described earlier, could be a safe vehicle for future research.

(3) There is a tendency to implement hybrid classifiers and ensembles for emotion classification, since there is a plethora of single classifiers that have been exhaustively assessed, without consistent results. It is also crucial to validate results in benchmarks that include samples from multiple databases and datasets.
(4) As far as databases are concerned, the complexity of the problem along with cross cultural diversities makes the development of a complete common database an extremely difficult task.

References

Aigner M, Sachs G, Bruckmüller E, Winklbaur B, Zitterl W, Kryspin-Exner I, Gur R, Katschnig H (2007) Cognitive and emotion recognition deficits in obsessive–compulsive disorder. Psychiatr Res 149:121–128

Anagnostopoulos CN, Iliou T (2010) Towards emotion recognition from speech: definition, problems and the materials of research. Stud Comput Intell 279:127–143

Anagnostopoulos CN, Vovoli E (2010) Sound processing features for speaker-dependent and phrase-independent emotion recognition in Berlin Database. In: Papadopoulos GA, Wojtkowski W, Wojtkowski G, Wrycza S, Zupancic J (eds) Information systems development, pp 413–421

Ang J, Dhillon R, Shriberg E, Stolcke A (2002) Prosody-based automatic detection of annoyance and frustration in human–computer dialog. In: Proceedings of interspeech, pp 2037–2040

Atassi H, Esposito A (2008) A speaker independent approach to the classification of emotional vocal expressions. In: Proceedings of 20th IEEE international conference on tools with artificial intelligence, pp 147–152

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

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

Batliner A, Steidl S, Schuller B, Seppi D, Laskowski K, Vogt T, Devillers L, Vidrascu L, Amir N, Kessous L, Aharonson V (2006) Combining efforts for improving automatic classification of emotional user states. In: Proceedings of 1st international language technologies conference, pp 240–245

Bogert B, Healy M, Tukey J (1963) The quefrency analysis of time series for echoes: cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking. In: Rosenblatt M (ed) Symposium on time series analysis. Wiley, New York, pp 209–243

Calder J, Lawrence AD, Young AW (2001) Neuropsychology of fear and loathing. Nat Rev Neurosci 2:352–363

Cheng XM, Cheng PY, Zhao L (2009) A study on emotional feature analysis and recognition in speech signal. In: Proceedings of international conference on measuring technology and mechatronics automation, pp 418–420

Cen L, Ser W, Yu ZL (2008) Speech emotion recognition using canonical correlation analysis and probabilistic neural network. In: Proceedings of 7th international conference on machine learning and applications, pp 859–862

Cowie R, Douglas-Cowie E, Savvidou S, McMahon E, Sawey M, Schroder M (2000) FEELTRACE: an instrument for recording perceived emotion in real time. In: Proceedings of ISCA speech and emotion workshop, pp 19–24

Cowie R, Douglas-Cowie E, Cox C (2005) Beyond emotion archetypes: databases for emotion modelling using neural networks. Neural Netw 18:371–388

Devillers L, Vasilescu I, Lamel L (2003) Emotion detection in task oriented spoken dialogs. In: Proceedings of IEEE multimedia human–machine interface and interaction conference, pp 549–552

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

Douglas-Cowie E, Cowie R, Sneddon I, Cox C, Lowry O, McRorie M, Martin JC, Devillers L, Abrilan S, Batliner A, Amir N, Karpouzis K (2007) The HUMAINE database: addressing the collection and annotation of naturalistic and induced emotional data. In: Proceedings of international conference affective computing and intelligent interaction, pp 488–500

Dumouche P, Dehak N, Attabi Y, Dehak R, Boufaden N (2009) Cepstral and long-term features for emotion recognition. In: Proceedings of INTERSPEECH, pp 344–347

Fernandez R, Picard RW (2003) Modeling drivers' speech under stress. Speech Communications, vol 40. Elsevier, pp 145–159

Firoz Shah A, Vimal Krishnan VR, Raji Sukumar A, Jayakumar A, Babu Anto P (2009) Speaker independent automatic emotion recognition from speech: a comparison of MFCCs and discrete wavelet transforms. In: Proceedings of international conference on advances in recent technologies in communication and computing, pp 528–531

Fontaine JRJ, Scherer KR, Roesch EB, Ellsworth PC (2010) The world of emotions is not two dimensional. Psychol Sci 18:1050–1057

Forbes-Riley K, Litman DJ (2004) Predicting emotion in spoken dialogue from multiple knowledge sources. In: Proceedings of human language technology conference, North American chapter of the association computational linguistics (HLT/NAACL), pp 201–208

France DJ, Shivavi RG, Silverman S, Silverman M, Wilkes M (2000) Acoustical properties of speech as indicators of depression and suicidal risk. IEEE Trans Biomed Eng, 7:829–837

Fu L, Mao X, Chen L (2008a) Relative speech emotion recognition based artificial neural network. In: Proceedings of IEEE Pacific-Asia workshop on computational intelligence and industrial application, pp 140–144

Fu L, Mao X, Chen L (2008b) Speaker independent emotion recognition using HMMs fusion system with relative features. In: Proceedings of 1st international conference on intelligent networks and intelligent systems, pp 608–611

Giannakopoulos T, Pikrakis A, Theodoridis S (2009) A dimensional approach to emotion recognition of speech from movies. In: Proceedings of IEEE international conference on acoustics, speech and signal processing, pp 65–68

Graciarena M, Shriberg E, Stolcke A, Enos F, Hirschberg J, Kajarekar S (2006) Combining prosodic lexical and cepstral systems for deceptive speech detection. In: Proceedings of IEEE international conference on acoustics, speech and signal processing, pp 1033–1036

Hanjalic A (2006) Extracting moods from pictures and sounds: towards truly personalized TV. IEEE Signal Process Mag 23:90–100

Hanjalic A, Xu LQ (2005) Affective video content representation and modeling. IEEE Trans Multimed 7: 143–154

Hoch S, Althoff F, McGlaun G, Rigoll G (2005) Bimodal fusion of emotional data in an automotive environment. In: Proceedings of international conference audio. Speech and Signal Processing, vol 2, pp 1085–1088

Hozjan V, Kacic Z (2006) Context-independent multilingual emotion recognition from speech signals. Int J Speech Technol 6:311–320

Ijima Y, Tachibana M, Nose T, Kobayashi T (2009) Emotional speech recognition based on style estimation and adaptation with multiple-regression HMM. In: Proceedings of 2009 IEEE international conference on acoustics, speech and signal processing, pp 4157–4160

Iliou T, Anagnostopoulos C-N (2009) Comparison of different classifiers for emotion recognition. In: Proceedings of panhellenic conference in informatics, pp 102–106

Jin Y, Zhao Y, Huang C, Zhao L (2009) Study on the emotion recognition of whispered speech. In: Proceedings of global congress on intelligent systems, pp 242–246

Kockmann M, Burget L, Cernocky J (2009) Brno university of technology system for interspeech 2009 emotion challenge. In: Proceedings of INTERSPEECH, pp 348–351

Kostoulas TP, Fakotakis N (2006) A speaker dependent emotion recognition framework, CSNDSP. In: Proceedings of 5th international symposium computers, systems, networks and digital signal processing, pp 305–309

Kostoulas T, Ganchev T, Mporas I, Fakotakis N (2007) Detection of negative emotional states in real-world scenario. In: Proceedings of 19th IEEE international conference on tools with artificial intelligence, pp 502–509

Kostoulas T, Ganchev T, Lazaridis A, Fakotakis N (2010) Enhancing Emotion recognition from speech through feature selection. In: Sojka P, Horák A, Kopecek I, Pala K (eds) Text, speech and dialogue, lecture notes in artificial intelligence, vol 6231, pp 338–344

Kwon OW, Chan K, Hao J, Lee TW (2003) Emotion recognition by speech signals. In: Proceedings of Eurospeech conference, pp 125–128

Lee CM, Narayanan SS (2005) Toward detecting emotions in spoken dialogs. IEEE Trans Speech Audio Process 13:293–303

Lee CM, Narayanan SS, Pieraccini R (2002) Combining acoustic and language information for emotion recognition. In: Proceedings of interspeech, pp 873–376

Lee CM, Yildirim S, Bulut M, Kazemzadeh A, Busso C, Deng Z, Lee SS, Narayanan S (2004) Emotion recognition based on phoneme classes. In: Proceedings of international conference spoken language processing, pp 205–211

Lee C, Mower E, Busso C, Lee S, Narayanan S (2009) Emotion recognition using a hierarchical binary decision tree approach. In: Proceedings of INTERSPEECH, pp 320–323
Litman DJ, Forbes-Riley K (2004) Predicting student emotions in computer-human tutoring dialogues In: Proceedings of 42nd annual meeting on association for computational linguistics
Luengo I, Navas E, Hernaez I (2010) Feature analysis and evaluation for automatic emotion identification in speech. IEEE Trans Multimed 12:490–501
Lugger M, Yang B (2007a) An incremental analysis of different feature groups in speaker independent emotion recognition. In: Proceedings of international congress phonetic sciences, pp 2149–2152
Lugger M, Yang B (2007b) The relevance of voice quality features in speaker independent emotion recognition. In: Proceedings of IEEE international conference on acoustics, speech and signal processing, pp 17–20
Manning CD, Schütze H (1999) Foundations of statistical natural language processing. MIT Press, Cambridge
Mao X, Chen L, Fu L (2009) Multi-level speech emotion recognition based on HMM and ANN. In: Proceedings of world congress on computer science and information engineering, pp 225–229
Matos S, Birring SS, Pavord ID, Evans DH (2006) Detection of cough signals in continuous audio recordings Using HMM. IEEE Trans Biomed Eng 53:1078–1083
Mishra HK, Sekhar CC (2009) Variational gaussian mixture models for speech emotion recognition. In: Proceedings of 7th international conference on advances in pattern recognition, pp 183–186
Morrison D, Wang R, Silva LCD (2007) Ensemble methods for spoken emotion recognition in call-centres. Speech Commun 49:98–112
Navas E, Hernáez I, Luengo I (2006) An objective and subjective study of the role of semantics and prosodic features in building corpora for emotional TTS. IEEE Trans Audio Speech Lang Process 14:1117–1127
Neiberg D, Elenius K, Laskowski K (2006) Emotion recognition in spontaneous speech using GMMs. In: Proceedings of INTERSPEECH conference, pp 809–812
Nogueiras A, Moreno A, Bonafonte A, Mariño JB (2001) Speech emotion recognition using Hidden Markov models. In: Proceedings of EUROSPEECH, pp 2679–2682
Nwe TL, Foo SW, De Silva LC (2003) Classification of stress in speech using linear and nonlinear features. In: Proceedings of IEEE international conference acoustics, speech, and signal processing, pp 9–12
Ortony A, Clore G, Collins A (1988) The cognitive structure of emotions. Cambridge University Press, Cambridge
Pal P, Iyer AN, Yantorno RE (2006) Emotion detection from infant facial expressions and cries. In: Proceedings of IEEE international conference on acoustics, speech, and signal processing, pp 721–724
Pao TL, Liao WY, Chen YT, Yeh JH, Cheng YM, Chien CS (2007a) Comparison of several classifiers for emotion recognition from noisy mandarin speech. In: Proceedings of 3rd international conference on international information hiding and multimedia signal processing, pp 23–26
Pao TL, Chien CS, Chen YT, Yeh JH, Cheng YM, Liao WY (2007b) Combination of multiple classifiers for improving emotion recognition in Mandarin speech. In: Proceedings of 3rd international conference on international information hiding and multimedia signal processing, pp 35–38
Petridis S, Pantic M (2008) Audiovisual discrimination between laughter and speech. In: Proceedings of IEEE international conference on acoustics, speech, and signal processing, pp 5117–5120
Rong J, Chen YPP, Chowdhury M, Li G (2007) Acoustic features extraction for emotion recognition. In: Proceedings 6th IEEE/ACIS international conference on computer and information science, pp 419–424
Russell JA, Weiss A, Mendelsohn GA (1989) Affect Grid: a single-item scale of pleasure and arousal. J Pers Soc Psychol 57:493–502
Russell JA, Bachorowski J, Fernandez-Dols J (2003) Facial and vocal expressions of emotion. Annu Revis Psychol 54:329–349
Schroder M (2003) Experimental study of affect bursts. Speech Commun 40:99–116
Schuller B, Rigoll G (2009) Recognising interest in conversational speech–comparing bag of frames and supra-segmental features. In: Proceedings of INTERSPEECH, pp 1999–2002
Schuller B, Rigoll G, Lang M (2003) Hidden Markov model-based speech emotion recognition. In: Proceedings of international conference on multimedia and expo, pp 401–404
Schuller B, Rigoll G, Lang M (2004) Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In: Proceedings of IEEE international conference acoustics, speech, and signal processing, pp. 577–580
Schuller B, Muller R, Lang M, Rigoll G (2005a) Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles. In: Proceedings of 9th Eurospeech–Interspeech, pp 805–809
Schuller B, Villar RJ, Rigoll G, Lang M (2005b) Meta-classifiers in acoustic and linguistic feature fusion-based affect recognition. In: Proceedings of IEEE international conference on acoustics, speech and signal processing, pp 325–328

Schuller B, Reiter S, Mueller R, Al-Hames M, Lang M, Rigoll G (2005c) Speaker-independent speech emotion recognition by ensemble classification. In: Proceedings international conference on multimedia and expo, pp 864–867

Schuller B, Reiter S, Rigoll G (2006) Evolutionary feature generation in speech emotion recognition. In: Proceedings 2006 IEEE international conference on multimedia and expo, pp 5–8

Schuller B, Batliner A, Seppi D, Steidl S, Vogt T, Wagner J, Devillers L, Vidrascu L, Amir N, Kessous L, Aharonson V (2007) The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals. In: Proceedings of INTERSPEECH, pp 2253–2256

Schuller B, Müller R, Eyben F, Gast J, Hörnler B, Wöllmer M, Rigoll G, Höthker A, Konosu H (2009) Being bored? Recognising natural interest by extensive audiovisual integration for real-life application. Image Vis Comput 27:1760–1774

Schuller B, Wollmer M, Eyben F, Rigoll G (2009) The role of prosody in affective speech. Peter Lan Publishing Group, Bern

Schuller B, Batliner A, Steidl S, Seppi D (2009c) Emotion recognition from speech: putting ASR in the loop. In: Proceedings of IEEE international conference on acoustics, speech and signal processing, pp 4585–4588

Schuller B, Schenk J, Rigoll G, Knaup T (2009d) "The Godfather" vs. "Chaos": comparing linguistic analysis based on on-line knowledge sources and bags-of-n-grams for movie review valence estimation. In: Proceedings of 10th international conference on document analysis and recognition, pp 858-862

Schuller B, Steidl S, Batliner A (2009e) The INTERSPEECH 2009 emotion challenge. In: Proceedings of INTERSPEECH, pp 312–315

Schuller B, Vlasenko B, Eyben F, Wollmer M, Stuhlsatz A, Wendemuth A, Rigoll G (2010) Cross-corpus acoustic emotion recognition: variances and strategies. IEEE Trans Affect Comput 1:119–131

Schuller B, Batliner A, Steidl S, Seppi D (2011) Recognising realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge. Speech Commun 53:1062–1087

Shami MT, Kamel MS (2005) Segment-based approach to the recognition of emotions in speech. In: Proceedings of IEEE international conference on multimedia and expo, pp 4–7

Stuhlsatz A, Meyer C, Eyben F, Zielke T, Meier G, Schuller B (2011) Deep neural networks for acoustic emotion recognition: raising the benchmarks. In: Proceedings international conference on acoustics speech and signal processing, pp 5688–5691

Sidorova J (2007) Speech emotion recognition. Ph.D. Thesis, Universitat Pompeu Fabra, Barcelona

Vlasenko B, Schuller B, Wendemut A, Rigoll G, Frame vs (2007) Turn-level: emotion recognition from speech considering static and dynamic processing. In: Proceedings 2nd international conference on affective computing and intelligent interaction, pp 139–147

Vogt T, André E (2005) Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. In: Proceedings IEEE international conference on multimedia and expo, pp 474–477

Vogt T, André E (2006) Improving automatic emotion recognition from speech via gender differentiation. In: Proceedings of language resources and evaluation conference, pp 1123–1126

Vogt T, André E (2009) Exploring the benefits of discretization of acoustic features for speech emotion recognition. In: Proceedings 10th INTERSPEECH conference, pp 328–331

Wagner J, Kim NJ, Andre E (2005) From physiological signals to emotions: implementing and comparing selected methods for feature extraction and classification. In: Proceedings of IEEE international conference multimedia and expo, pp 940–943

Wang Y, Du S, Zhan Y (2008) Adaptive and optimal classification of speech emotion recognition. In: Proceedings of 4th international conference on natural computation, pp 407–411

Wang S, Ling X, Zhang F, Tong J (2010) Speech emotion recognition based on principal component analysis and back propagation neural network. In: Proceedings of international conference on measuring technology and mechatronics automation, pp 437–440

Wenjing H, Haifeng L, Chunyu G (2009) A hybrid speech emotion perception method of VQ-based feature processing and ANN recognition. In: Proceedings of global congress on intelligent systems, pp 145–149

Wierzbicka A (1999) Emotions across languages and cultures: diversity and universals. Cambridge University Press, Cambridge

Wu CH, Liang WB (2011) Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. IEEE Trans Affect Comput 2:10–21

Wu CH, Chuang ZJ, Lin YC (2006) Emotion recognition from text using semantic label and separable mixture model. ACM Trans Asian Lang Inf Process 5:165–182

Wu S, Falk TH, Chan WY (2009) Automatic recognition of speech emotion using long-term spectro-temporal features. In: Proceedings of 16th international conference on digital signal processing

Yang C, Ji L, Liu G (2009a) Study to speech emotion recognition based on TWINsSVM. In: Proceedings of 5th international conference on natural computation, pp 312–316

Yang T, Yang J, Bi F (2009b) Emotion statuses recognition of speech signal using intuitionistic fuzzy set. In: Proceedings of world congress on software engineering, pp 204–207

You M, Chen C, Bu J, Liu J, Tao J (2006) Emotional speech analysis on nonlinear manifold. In: Proceedings of 18th international conference on pattern recognition, pp 91–94

Yu W (2008) Research and implementation of emotional feature classification and recognition in speech signal. In: Proceedings of international symposium on intelligent information technology application, pp 471–474

Yun S, Yoo CD, (2009) Speech emotion recognition via a max-margin framework incorporating a loss function based on the Watson and Tellegen's emotion model. In: Proceedings IEEE international conference on acoustics, speech and signal processing, pp 4169–4172

Zeng Z, Pantic M, Roisman GI, Huang TS (2009) A survey of affect recognition methods: audio, visual, and spontaneous expressions. IEEE Trans Pattern Anal Mach Intell 31:39–58

Zhou Y, Zhang J, Wang L, Yan Y (2009a) Emotion recognition and conversion for mandarin speech. In: Proceedings of 6th international conference on fuzzy systems and knowledge discovery, pp 179–183

Zhou Y, Sun Y, Yang L, Yan Y (2009b) Applying articulatory features to speech emotion recognition. In: Proceedings of international conference on research challenges in computer science, pp 73–76