In *Social emotions in nature and artifact: emotions in human and human-computer interaction*, S. Marsella J. Gratch, Ed. Oxford University Press, New York, NY, USA, 2012 (press)

Toward effective automatic recognition systems of emotion in speech

Carlos Busso[2], Murtaza Bulut[3], and Shrikanth Narayanan[1]

[1] Viterbi School of Engineering, University of Southern California, Los Angeles, CA 90089, USA
shri@sipi.usc.edu
[2] The University of Texas at Dallas, Richardson TX 75080, USA
busso@utdallas.edu
[3] Philips Research, Eindhoven, Netherlands
murtaza.bulut@philips.com

1 Introduction

Humans are emotional beings and emotions are one of the main drivers of human thoughts and actions. Therefore, for all environments designed for humans, it is essential that emotion processing capabilities such as analysis, recognition, and synthesis are incorporated. Naturally, any type of information, such as audio, visual, written, mental or physiological, can be used for these tasks.

In this chapter, our concentration will be on emotion recognition from speech. Specifically, this chapter discusses the collection and organization of databases and emotional descriptors, the calculation, selection, and normalization of relevant speech features, and the models used to recognize emotions. We outline achievements, open questions, and future challenges in building Effective Automatic Speech Emotion Recognition (EASER) systems.

It is known that emotions cause mental and physiological changes which also reflect in uttered speech. When processing the generated speech, one can calculate different features, which can be utilized to learn the relationship between features and emotions. Once such relationship are learned, theoretically, one can calculate the features and then automatically recognize the emotions present in speech.

From a scientific perspective, recognition of emotions is nothing more than a mapping from a feature space to emotion descriptors or labels space. For the mapping between the two spaces, different machine learning algorithms have been used [30]. In general, theories to perform the mapping have solid analytical foundations and are well defined and validated. Hardly, however, is the same true for feature and emotion spaces. In other words, it is a challenging issue to determine which features to use and how to describe emotions. The

problem of emotion recognition from speech critically depends on these two factors, meaning that high and robust emotion recognition performance can be achieved only with the accurate selection of features and emotional labels. In this chapter, selection of correct features and emotional labels will be discussed in the view of building EASER systems.

For emotion recognition from speech, time, frequency, and lexical features have been popularly used [23, 44]. Examples of time domain features are fundamental frequency, duration, and energy statistics. Frequency domain features are found by applying a transform to the time domain signal. Examples of such features are Mel Frequency Cepstral Coefficients (MFCCs), Mel Filter Bank (MFB) [12] coefficients, or other perceptually motivated features [39]. Lexical features, although not as popularly used as time and frequency domain features, can also be very effective as shown in [46].

There have been many studies of how to classify and describe emotions. Two of the most popular approaches are to use categorical labels or dimensional labels. Examples of categorical labels are anger, sadness, disgust, and happiness. Considering the large number of possible categorical labels - for robust and effective user and application specific emotion recognition applications - it is critical to select the most appropriate emotional labels and to train the emotion recognizer accordingly. When dimensional labels, such as activation, valence, and dominance, are used, emotions are described with continuous values. Later, the dimensional labels can be mapped to categorical emotional labels, if needed [36]. Each technique has its own advantages and disadvantages. As discussed in the following sections, for robust EASER systems - that can adapt to different users, environments, and applications - it is important to use the correct emotional labels.

Although it is not popularly used in any real life applications today, we expect that recognizing emotions from speech will be very important and popular in future applications and products. Examples of possible application areas are automated call centers where a caller can be forwarded to a human agent when particular emotions are detected. In ambient intelligent (AmI) environments, atmosphere and devices can be adapted according to users' emotions. In future classrooms, students' emotional states can be used as feedback for teachers. In automated meeting summary generation and speaker diarization systems, emotional information can be added. In short, we can expect that having information about users' emotional states will improve the human-machine interactions (HMI) and therefore increase the productivity and satisfaction of the users.

The rest of the chapter is organized as follows. Section 2 gives a brief overview on the state of the art in emotion recognition from speech. Section 3 discusses in detail some aspects that need to be carefully considered for EASER systems; it also describes our contribution in the field. Section 4 presents our perspective on the directions that this area needs to take. Finally, Section 5 summarizes the chapter.

2 Overview

The idea of recognizing emotions in speech has been of interest for many years. A quick search will produce many scientific publications on the topic. It is out of the scope of this chapter to present a detailed review of the existing literature. Instead, our focus is on selected aspects that are crucial for a robust speech emotion recognition system. The reader can refer to [23] and [73] for reviews.

As in any pattern recognition problem, the performance of emotion recognition from speech depends on type, organization, and representation of training data. It is obvious, but nevertheless useful to recall, that a good training dataset is the one that performs well for a targeted test set or application. This statement has the following implication. *There is no training dataset that will perform well in all conditions.* Therefore, it is important that data collection, organization, and labeling are performed by taking into account the target application and users. The reverse approach of first gathering data and then defining specifications of application and target users is not suggested, as it will hardly reach the optimal performance. At least, not for a task as challenging as recognizing emotions from speech.

A popular approach to recognizing emotions is based on using acoustical information [63], such as prosody and spectral features. The number of features used varies depending on the application. Having a large number of features increases the complexity of the system and normally results in longer system training times. Therefore, a popular approach is to start with a larger set of features and then eliminate the less significant features to generate a more compact and robust feature set. As expected, the final compact feature sets can vary based on the database. This means that for different speakers' emotions and conditions, different feature sets can perform better. See section 3.3 for a detailed discussion on features.

In addition to the acoustic features, lexical, semantic, and discourse information can also be used. In [44], it is shown that using language information in addition to acoustic information improves emotion recognition rates significantly.

Various recognition methods have been used in the literature. Popularly used machine learning techniques [30] are linear discriminators, Gaussian Mixture Models (GMMs), Hidden Markov Models (HMMs), neural networks (NNs), Bayes classifiers, and fuzzy classifiers. In general, any of these techniques can be employed and be advantageous to others in certain conditions, depending on signal to noise (SNR) ratio, recorded emotions, recording conditions, and type and size of training databases. For example, in [53], k-nearest neighbors (KNN), GMM, HMM, weighted categorical average patterns (WCAP), and weighted discrete-KNN (W-DKNN) pattern recognition methods are compared for recognizing anger, happiness, sadness, neural state, and boredom from noisy Mandarin emotional speech, and it was found that W-DKNN performs the best.

Humans take into account all of the factors, including visual, vocal, cultural, environmental, and personal clues, to classify emotions [57, 56]. Although it can vary significantly, in general, human performance for recognizing emotions from vocal cues is around 80%. This value can be considered as a reasonable upper limit performance that should be expected from speech emotion recognizers. As expected, some emotions are easier to recognize than others. For example, humans are much better at recognizing anger than recognizing happiness; so are the machines.

3 Analysis of effective automatic speech emotion recognition (EASER) systems' components

This section discusses essential aspects in the design and implementation of EASER systems. The section also presents some of our own contributions to the field.

3.1 Databases

The machine learning algorithms and statistical models used in emotion recognition systems are trained and tested with data that describes the problem at hand (data-driven approach). Therefore, the quality of emotional databases is extremely important.

Actors have played an important role in studying emotions [27]. The main advantage of recording acted databases is that many aspects of the recording can be carefully and systematically controlled (e.g., location of the microphones/cameras, emotional and lexical content, and noise free environment). Unfortunately, the elicitation techniques used for this task were not in accord with the well-established theories and methods used in the field of theater. In fact, most of the emotional databases in early work were recorded from actors or naïve subjects without acting experience who were asked to read sentences expressing given emotions (e.g., "read this sentence with this emotion"). As a result, the actors needed to cope with settings that were not natural for expressing emotions. The samples collected with this approach were characterized by highly prototyped emotions, which usually differed from the emotions displayed by regular people in real life situations in which mixtures of emotions are found [29, 26]. In real life situations, emotions may or may not be exaggerated and usually they consist of mixtures of several emotions. As a result, the models trained with data collected in laboratory settings are not easy to apply in real-life applications [5]. We have argued that the main problem of acted databases is not the use of actors but the methodology used to elicit the emotions [15, 10]. As suggested by the appraisal theory, emotions are expressed as a reaction to events (appraisal theory [47, 35]). Therefore, instead of giving a specific emotion to the actors, the researchers should give specific scenarios that will trigger the target emotions. For example, we collected the

interactive emotional dyadic motion capture (IEMOCAP) database, in which we used emotionally rich scripts and improvisations of fictitious situations, which were carefully designed to elicit specific emotions [10].These two elicitation techniques are rooted in the theatrical performance and are familiar to trained actors [31, 16].

Recent efforts on recording emotional databases have been focused on natural databases (i.e., non-acted emotions). Broadcast television programs have been extensively used for this purpose (VAM [37], EmoTV1 [1], Belfast naturalistic database [27]) . Other interesting approaches were based on recording in situ (Genova Airport Lost Luggage database [58]), Wizard of Oz interfaces (SmartKom [59], FAU AIBO [67]), interviews (AVIC [62], SAL [21]) and call center costumer care (CCD [44], CEMO [71]). Despite limitations such as copyright issues and lack of control, these databases are an important step forward in the area of automatic emotion recognition. Given the multiple variables considered in the study of emotions, it is clear that a collection of different databases rather than a single corpus will be needed to address many of the open questions in this multidisciplinary area. The HUMAINE project portal presents further descriptions of some of the existing emotional databases [38].

Research areas such as music retrieval (Music Information Retrieval Evaluation eXchange (MIREX) [49]) and different spoken language technologies (NIST [51]) have greatly benefited from having open evaluations in which different approaches are compared under similar conditions. In this direction, similarly for emotion recognition, seven research centers participated in the *combining efforts for improving automatic classification of emotional user state* (CEICES) initiative [7]. The task was later extended to the research community in the InterSpeech 2009 Emotion Challenge [64]. In these competitions, the FAU AIBO corpus [67] was used. This database was recorded from German children (10-13 years) who verbally interacted with a robot controlled by a human. Building upon these initiatives, it will be beneficial to add new databases for benchmark tests to include other sources of variability such as age, recording conditions, modalities, and languages. Fortunately, the recent trend for the new emotional databases is to make them available (e.g., VAM, Belfast naturalistic database, SAL) [28].

3.2 Emotional descriptors

Scherer proposed using an adapted version of the Brunswik's lens model to study vocal communication of the emotions [56]. This model makes an explicit distinction between the encoding (speaker), the transmission, and the representation (listener) of the emotion. The speaker encodes his/her emotional state in the speech (and other modalities) producing *distal indicators* that are transmitted. The listener perceives the information, referred to as *proximal cues* in the models, and makes inferences about their attributes. All these distinctions in the models are made because expression and perception

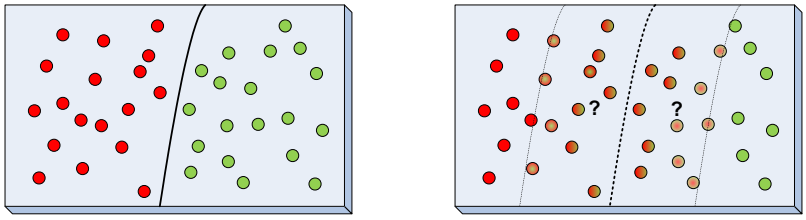
are two distinct and complex problems. The intended emotion encoded by the speaker may not necessarily match with the perceived emotion [14]. The distal indicators may be different from the proximal indicators (e.g., distortion in the transmission, structural characteristic of the perceptual organ) [56]. The process that transforms proximal cues into emotional attributes is intrinsically speaker dependent. As a result, it is not surprising that representing emotions is one of the most challenging tasks in emotion recognition.

Two of the most common strategies to characterize emotions are discrete categorical labels and continuous primitives attributes [55, 20]. With discrete labels, the emotional databases are evaluated in terms of words such as anger, happiness, and sadness. With continuous attributes, the emotional content of the databases is projected into different dimensions with emotional connotation. The most used attributes/dimensions are valence (measuring how positive or negative the subject is) and activation or arousal (how active or passive the subject is). A third dimension, such as dominance or control, is sometimes included to make a distinction between certain emotions that share similar activation-valence properties (e.g., fear and anger). Both representations have advantages and disadvantages. For example, inter-speaker agreement is usually higher with continuous attributes. However, categorical descriptors simplify the design of interfaces [55]. We believe that both approaches provide useful complementary information to describe the emotional content of the speaker. For instance, continuous attributes are useful to differentiate intensity levels within samples labeled with the same emotional class.

Regardless of the approach used to represent emotions, the real emotional labels or values of the attributes are unknown. As an approximation, subjective perceptive evaluations have been commonly used. These assessments are expensive and time consuming. Therefore, a limited number of labelers assess the emotional content of each sample in the corpus (e.g., [database - number of evaluators] IEMOCAP-3, VAM-17, AVIC-4, FAU AIBO-5, CCD-4). Since these evaluations are usually characterized by low inter-speaker agreement, the final tags assigned to the samples are inherently noisy. This is clearly observed with non-pure emotions frequently observed in real-life interactions [26]. We have studied the mismatch between the expression and perception of the emotions [14]. Our results suggested that tags assigned by labelers might not accurately describe the true emotions conveyed by speakers. These results agree with the work of Biersack and Kempe [8]. They conducted a study with 200 speakers and 4 groups of 20 listeners on speech of one-minute average duration. The study showed that the happiness mood rating reported by the speakers (i.e., self rating) was not correlated to the happiness rating perceived by listeners (i.e, observers). This was viewed as an indication that other factors besides vocal cues also play an important role in emotion perception. Another study that investigated observer and self-annotation differences and similarities was done by recording vocal and facial expressions during a multiplayer video game [69]. The emotion ratings were done in arousal (i.e.,

how active or passive) and valence (i.e., how positive or negative) dimensions on a scale from 0 to 100 for audio only, video only, audio and visual, and audio and visual plus content data. It was found that self-rating can significantly differ from observer ratings. As expected, agreements on valence and arousal dimension were also different from each other. Since the ultimate goal of an emotion recognition system is not to recognize what others perceive, but what the user expresses or feels, subjective experiments should be viewed as an approximation.

The confusion in the emotional labels is one of the main differences between emotion recognition and conventional machine learning problems (Fig. 1). The underlying classes in a recognition problem are perfectly defined, even when no hyperplane perfectly separates the groups. If we are interested in recognizing emotion, this property is far from true, and the models have to cope with this variability. It is for this reason that some researchers have stated that emotion recognition is an ill-defined problem.



(a) Simple machine learning task (b) Emotion recognition task

Fig. 1. Main challenge in emotion recognition is the lack of clear emotional labels. Given the important differences in perception, the boundaries in emotional categories are blurred and the models have to cope with this variability.

If discrete categorical labels are used, the emotional classes need to be defined. In general, there is a tradeoff between inter-evaluator agreement and description accuracy. If the number of emotion categories is too extensive, the agreement between evaluators will be low. If the list of emotional classes is limited, the emotional description of the utterances will be poor and likely less accurate. One popular approach is to use large numbers of classes, which are later clustered in broad emotional classes. For example, the FAU AIBO database was originally evaluated with the labels joyful, surprised, emphatic, helpless, touchy, angry, motherese, bored, reprimanding, neutral, and others [64]. These classes were grouped into 5 general categories (anger, emphatic, neutral, positive and other). Another example is the SAFE corpus, in which over 20 emotional labels were grouped into 4 broad categories (fear, other negative emotions, neutral and positive emotions) [19]. The main problem of this approach is identifying how to define the emotional partition without increasing the noise in the labeling. Instead of using ad-hoc methods, we have

proposed the use of an interval type-2 fuzzy logic system to map and cluster emotional terms [40]. In this approach, the emotional words are evaluated in terms of valence, activation, and dominance (VAD) (e.g., how do you perceive the concept "happiness" in the VAD space?). Instead of selecting a single value, double sliders are used to enclose the range in which the emotional labels are believed to be. Therefore, inter-subject uncertainty is directly included in the model. This information is used to create interval type-2 fuzzy sets, which are used to map one vocabulary to another. Likewise, if databases are labeled with different emotional categories, this approach can be used to translate the labels into a common vocabulary set.

It is also unclear what is the best time scale to evaluate emotions. The conventional approach is to tag sentences or turns. However, the emotional content may not be constant within the sentences, especially for long samples [6]. The FAU AIBO database was labeled at word level to address this problem [67]. A heuristic approach was later used to map the labels to longer units (sentences or chunks). However, with short time units labelers may not have enough evidence to make the assessment, decreasing the inter-evaluator agreement. Also, it has been shown that the performance of emotion recognition is lower for short speech segments [42]. An alternative approach was presented by Cowie *et al.* to continuously evaluate the emotional content of data using the tool FEELTRACE [22]. As labelers watch the clips, they are asked to continuously move a cursor controlled by a mouse in a valence-activation space. The 2D space is enriched with strategically located categorical labels. The samples do not need to be segmented for evaluation.

While evaluating a database, many variables need to be defined. As an example, consider a multi-modal database where both vocal and visual cues were recorded from actors performing short scripts (as in [10]). For such datasets, one can use vocal cues only, visual cues only, or vocal and visual cues together to label the perceived emotions. In addition, the emotional classification by listeners can be done on randomly distributed, isolated (i.e., out of context) samples. Listeners may receive a list of emotions to select one or more categories to describe the emotional content. Or, as an alternative, the evaluation can be completely open choice by asking the listeners to enter the emotion or emotions that they perceive. All these variations present advantages and disadvantages that need to be balanced in design.

We believe that the emotional labels need to be driven by the application at hand. In our previous work with the call center customer care database, the annotation was simplified to either negative or non-negative [44]. We have also argued that for many applications it may be enough to detect emotional speech (neutral versus emotional speech) [13]. By making the problem more concrete and specific, automatic emotion recognition system can become feasible tools.

Characteristics of human evaluators can also be very important, as one can expect differences in emotion perception due to differences in age, sex, culture, education, experience, and other personal factors. For a complete description of an emotional database, detailed profiles of evaluators should

also be included. Such descriptions will help in defining the user profiles for the automatic emotion recognition systems. Having the systems combined with user profiles will help to improve the usage and performance of the systems.

3.3 Features

As in any machine learning problem, features with discriminative power are important for emotion recognition. A summary with the features most used to recognize emotions is presented in Table 1. The list includes prosodic, spectral, and voice quality features. In addition to acoustic features, lexical and discourse features have also been proposed [25, 6, 44]. In fact, we have shown that these features are useful in the context of call center applications [44]. This section discusses only acoustic features.

Description	Features
Supra-segmental acoustic features (prosody)	- Pitch: mean, median, standard deviation, maximum, minimum, range (max-min), linear regression coefficient, lower and upper quartile, kurtosis, skewness, slope, curvature, inflection - Energy: mean, median, standard deviation, maximum, minimum, range, linear regression coefficient - Duration: speech-rate, ration of duration of voiced and unvoiced region, duration of longest voiced region - Zero crossing-rate
Segmental acoustic features (Short-term spectrum of speech)	- Mel-frequency cepstral coefficients (MFCC) - Mel filter bank (MFB) - Spectral centroid - Formant: F1, F2 and their bandwidth BW1, BW2
Voice quality features (intrasegmental level)	- Jitter (pitch modulation) - Shimmer (amplitude modulation) - Harmonics to Noise Ratio (HNR) - Noise-to-Harmonics Ratio (NHR) - Normalized amplitude quotient (NAQ)

Table 1. Common acoustic features used in emotion recognition (based on the following studies [44, 13, 19, 61, 66])

Different combinations of speech features have been proposed for emotion recognition. In machine-learning problems, the underlying conditional probability distributions are commonly unknown. Therefore, they have to be approximated from the test data. For a fixed number of training samples, the quality in the distribution approximation decreases as the dimensionality of the problem increases [32]. Therefore, non-relevant features will decrease the performance of the classifier. This problem, also known as the curse of dimensionality, is especially observed when non-parametrical distributions are assumed, in which more information is required from the data.

The standard approach in current emotion recognition systems is to compute a big feature vector containing all relevant acoustic information (in some cases higher than 4,000 [61]). Then, the feature vector is reduced to a subset

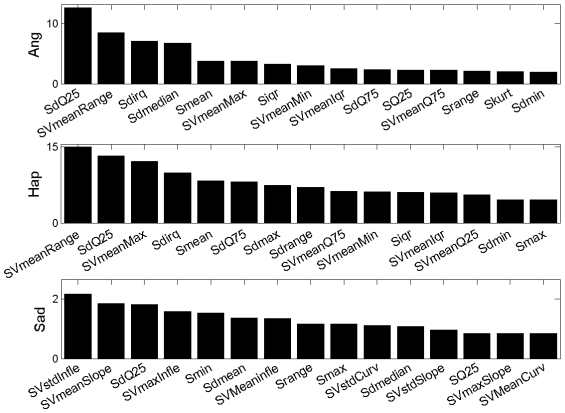
that provides better discrimination for the given task using feature selection techniques, such as forward or backward features selection, sequential forward floating search, genetic algorithms, evolutionary algorithms, linear discriminant analysis, principal component analysis, and information gain ratio [2, 70, 65, 60]. Clavel *et al.* proposed an interesting modification based on a two-step approach [19]. The acoustic features are separated in broad categories (spectral, prosodic, and voice quality features). In the first step, the best features within each group are selected. In the second step, the final feature set is selected from the candidate features. This approach is appealing since it enforces to some extent the contribution of features describing different aspects of speech. Even with this approach, the selected features are sensitive to the training and testing conditions (database, emotional descriptors, recording environment). Figure 2 shows the most emotionally salient statistics from the fundamental frequency for two databases (EMA [46] and EMO-DB, [9]). As expected, the figure shows that the ranking of the best features depends on the database and the emotional labels. These examples indicate that a robust emotion recognition system should use features that are found to convey emotional information across corpora.

As an alternative approach, we have proposed to study in detail the emotional modulation observed in acoustic features [13]. In the analysis, we compared different statistics derived from the fundamental frequency in terms of their emotional modulation. The distributions of pitch-related features extracted from emotional and neutral speech were compared using symmetric Kullback-Leibler distance. Then, the emotionally discriminative power of the pitch features was quantified by comparing nested logistic regression models. For generalization purpose, we considered cross-corpora tests with different emotional classes, speakers, recording settings, and languages. The results of this analysis suggested that gross pitch contour statistics such as mean, maximum, minimum, and range are more emotionally prominent than features describing the pitch shape, which may be dependent on the lexical content. In the final set of features used to detect emotional speech, the features were not necessarily the ones that maximize the performance for these databases, but the ones that in the analysis were found more emotionally prominent, according to the proposed experiments.

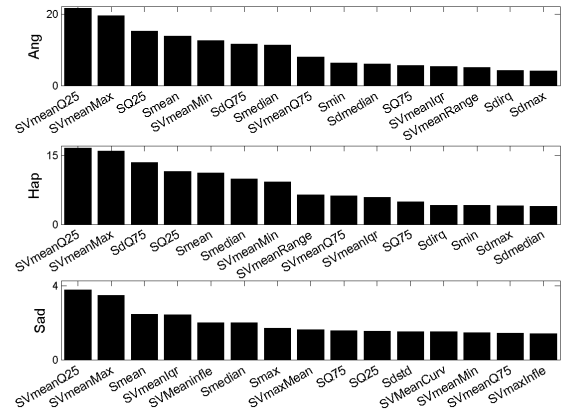
3.4 Data normalization

Data normalization is an important aspect that needs to be considered for a robust automatic emotion recognition system [42]. Ideally, the normalization step should remove or reduce all variability in sources, while preserving the emotional differences conveyed in the speech. Two of the most important sources of variability are recording conditions and inter-speaker variability.

The quality of the signal highly depends on the sensors used to capture the speech. Close-talking microphones (e.g., headphones) provide the best speech



(a) EMA database



(b) EMO-DB database

Fig. 2. Most emotionally prominent features from fundamental frequency. The figures were created by estimating the symmetric Kullback-Leibler Distance between the distribution of the features derived from neutral and emotional speech. The details are given in [13].

quality. However, they are not suitable for certain applications in which non-intrusive sensors are required (smart room or ambient intelligent environments). In those cases, the system may receive far-field reverberant speech with low signal-to-noise ratio (SNR). Likewise, if the speech is recorded using a phone or mobile speech system, the frequency bandwidth will be affected. The features derived from the speech signals will be directly affected by these distortions. In any of these cases, a robust emotion recognition system should be designed to attenuate possible mismatches between the speech set that was used to train the models and the speech set that is collected in the real-life applications. For instance, it is well-known that energy tends to increase with

angry or happy speech [20, 3, 50]. However, if the energy of the speech signal is not properly normalized, any difference in the microphone gain will affect the performance (i.e., loud speech may be confused with emotional speech).

Speech production is the result of controlled anatomical movements of the lungs, trachea, larynx, pharyngeal cavity, oral cavity, and nasal cavity. As a result, the properties of the speech are intrinsically speaker dependent. In fact, speech has been widely used for speaker identification [17]. Interestingly, some of the same features used for speaker recognition have also been proposed for emotion recognition. A robust emotion recognition system should compensate for speaker variability.

Let us consider, for example, the fundamental frequency mean, which has been extensively used as a feature to recognize emotion. In fact, our previous analysis indicated that the F0 mean is one of the most emotionally prominent aspects of the F0 contour, when properly normalized [13]. The fundamental frequency is directly constrained by the structure and size of the larynx [24]. While the F0 contour for men is bounded in the range 50-250 Hz, women can reach much higher F0 values (120-500 Hz) [24]. Figure 3 shows the distribution of the F0 mean in terms of gender, using the popular read-speech TIMIT database [33]. The F0 mean for each of the 630 subjects was estimated (one value for each subject). In addition, the figure shows data from 26 children (10-13 years) recorded in the training set of the FAU AIBO corpus [67, 64]. Three separate distributions for men, women, and children are clearly seen. As a result, any emotional difference will be blurred by inter-speaker differences. This point is clearly observed in Figure 4-a. In this figure, the F0 mean is computed at sentence level to estimate the distribution of anger and neutral speech across ten speakers recorded in the IEMOCAP database [10]. Although, in general, angry speech has higher F0 values than neutral speech, mixing emotional and speaker variations will result in noisy measures in the system.

Most of the current approaches to normalize speech or speech features are based on gross manipulation of the speech at sentence level. In many cases, the normalization approach is not clearly defined. Some of the approaches that have been widely used are Z-normalization (subtract the mean and divide by the standard deviation) [44], min-max normalization (scaling features between -1 and 1) [19], and subtraction of mean values [42]. For a given lexical unit (i.e., word or phoneme), Batliner *et al.* proposed to normalize speech features by estimating reference values for "average speakers" learned from a training database [4]. These reference values were used to scale the duration and energy of the speech.

We have proposed a two-step speaker dependent approach to normalize the energy and the fundamental frequency [13]. The main idea is to estimate the scaling parameter using only the neutral set of the emotional database. Assuming that the speaker's identity is known, the energy and pitch are linearly modified so that their mean values are equal to predefined reference values, estimated from the Wall Street Journal-based Continuous Speech Recognition Corpus Phase II [54] corpora. Then, the normalization parameters are applied

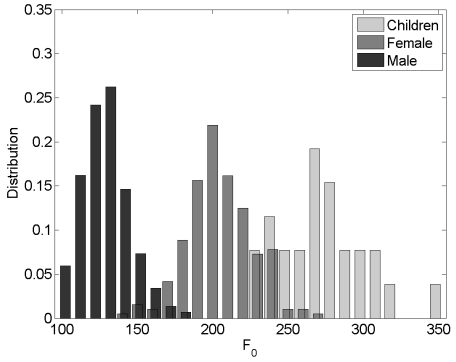
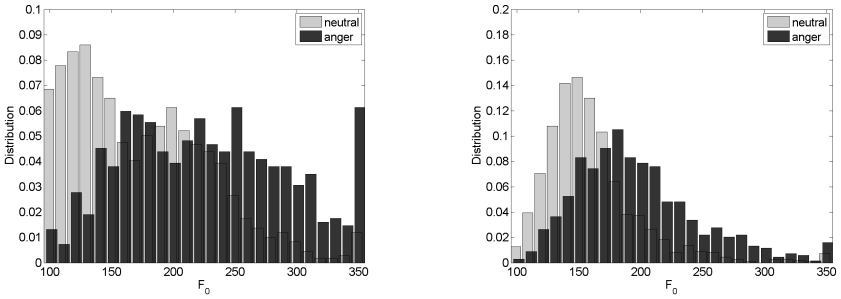


Fig. 3. Interspeaker variability in pitch mean (neutral speech)

to all speech samples from that speaker, including the emotional speech set. The scaling factors will not affect emotional discrimination in the speech, since the differences in the energy and the fundamental frequency contour across emotional categories will be preserved. Figure 4-b shows the distribution of neutral and angry speech in the IEMOCAP database after pitch normalization. Now, the shift in the distributions can be directly associated to emotional variations.



(a) F0 distribution (Gender) (b) F0 distribution (emotion)

Fig. 4. F0 mean computed at sentence level for angry and neutral speech (a) before normalization, and (b) after normalization. The figure shows that mixing emotional and speaker variations will result in noisy measures in the system.

One assumption made in this two-step approach is that neutral speech will be available for each speaker. The implications of this assumption are that speaker identities are known and that emotional labels for a portion of the data is known. For real-life applications, this assumption is reasonable when either the speakers are known or a few seconds of their neutral speech

can be pre-recorded. We are currently working on extending the proposed approach by using speaker-independent normalization. The first implication can be addressed with unsupervised speaker identification. The second implication can be addressed with reinforcement framework as displayed in Figure 5.

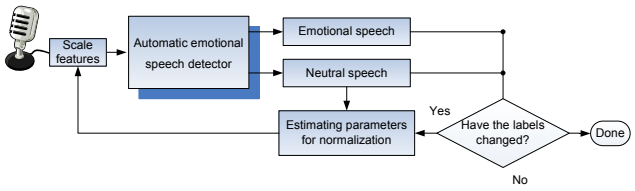


Fig. 5. Approach for speaker independent normalization. After the data is clustered using unsupervised speaker identification, an automatic emotional speech detector is used to identify neutral samples for a given speaker. The scaling factors are estimated from this neutral speech set. The process is repeated until the labels are no longer modified.

3.5 Models

In daily life, people express their emotions in an exaggerated manner only in certain conditions. Most of the time subtle emotions are expressed. Obviously, models trained on "highly emotional" data will perform well only in certain instances but poorly in general. It is still a challenging and open research area of how to process real life emotions, especially when only vocal data is present. Selection of database, emotional descriptors, normalization, and feature will have an effect on the performance and, therefore, on the architecture and models selected to build an emotional speech recognizer.

In previous works, variations of machine learning approaches have been proposed for emotion recognition. Some examples are support vector machines (SVMs) [45, 42, 64], Gaussian mixture models (GMMs) [13, 19], hidden Markov models (HMMs) [12, 45, 64], fuzzy logic estimators [36, 43], neural networks (NNs) [42, 6, 4], and linear discriminant classifiers (LDCs) [44, 42, 6]. These classifiers are usually divided into two categories: static and dynamic modeling. On the one hand, static classifiers use global features derived over the entire speech segment. They usually include statistics from supra-segmental acoustic features such as F0 range, mean duration, etc. (see Table 1). On the other hand, dynamic classifiers receive acoustic features at the frame level (i.e., 10-100 milliseconds). They capture the dynamic behavior of segmental acoustic features.

Instead of recognizing emotional classes, the system can be designed to estimate continuous values of the emotional primitives. We have used the

rule-based fuzzy logic estimator to infer the valence, activation, and dominance of the speech [36]. Using acted and spontaneous corpora, the estimations were found to be moderately to highly correlated with human evaluations ($0.42 < r < 0.85$). In addition, this representation can be used as a mid-level representation for categorical emotion recognition. Using k-nearest neighbor classifier, these attributes were mapped into emotional categories showing an overall recognition rate up to 83.5%.

In our previous work, we have addressed the simplified problem of detecting emotional speech [12, 13]. For this binary problem (which included neutral and emotional classes), we proposed the use of generic models trained with emotionally neutral reference speech (see Figure 6). The models are used to contrast the input speech. The underlying assumption is that expressive speech will differ from neutral speech in the feature space. Therefore, speech samples that differ in any aspect from neutral speech will not accurately fit the models. Therefore, a fitness measure such as the likelihood scores can be used as a feature to detect emotional speech. One advantage of this approach is that there are many neutral corpora available to train robust neutral models. These models do not depend on the emotional databases, speakers, or the emotional labels used to evaluate the speech. For the neutral models, we have proposed HMM for spectral features [12] and GMM for prosodic features [13]. In both cases, we used simple linear classifiers to discriminate between emotional and neutral speech using the likelihood scores as features. This framework not only performs better than a conventional emotion recognition system trained with the speech features but also generalizes better when there is a mismatch between the training and testing conditions [13].

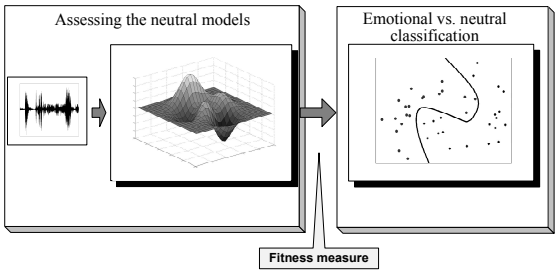


Fig. 6. General framework of the proposed two-step approach to discriminate neutral versus emotional speech. In the first step, the input speech is contrasted with robust neutral references models. In the second step, the *fitness measures* are used for binary emotional classification (details are given in [12, 13]).

In many applications, neutral speech is more common than expressive speech. For example, 60% of the FAU AIBO data is neutral speech in spite of the explicit elicitation techniques used to induce emotion in the children. Therefore, it will be useful to have a hierarchical approach in which a ro-

bust classifier is first used to detect emotional speech. Note that this step is independent of the application. Later, the emotional speech can be further processed using emotion specific models driven by the application at hand.

Using different machine learning frameworks, previous studies on automatic categorization of emotional speech have shown accuracy between 50% and 85% depending on the task (e.g. number of emotion labels, number of speakers, size of database) [52]. As expected, having a well defined training set with high agreement (i.e., "high prototypical") on emotional content among different listeners will perform better than a database with less agreement (i.e., "less prototypical") [66]. It is important to highlight that it is unfeasible and unrealistic to target performance near perfection. As mentioned in section 3.2, the perception of emotion is dependent on the listener. Therefore, emotional labels are noisy. In fact, Steidl *et al.* proposed to include the inherent inter-emotion confusion in the evaluation of emotion recognition performance [68]. If the system made errors similar to the human labelers, the performance was not considered completely wrong. Likewise, it is important to remember that speech is only one of the modalities that we use to express emotion. Even for humans, it is challenging to distinguish between certain emotions based only on speech. For example, we have shown that the fundamental frequency is a more informative measure for arousal of speech than valence of speech [11, 13]. Figure 7 shows the emotional classes in which the Nagelkerke r-square of logistic regression models between neutral and emotional speech (e.g., neutral versus anger) was higher (black) or lower (gray) than 0.5 (the location of the emotion in the activation-valence space was approximated from the FEELTRACE snapshots [23, 20]). This figure suggests that only emotion with high activation can be discriminated from neutral speech using fundamental frequency. Therefore, emotions like happiness and anger which differ in the valence dimension are usually confused [72]. This confusion is observed even when other acoustic features are used.

4 Future direction

The research in emotion recognition has progressed significantly in the past years. We expect further accelerated growth, especially when emotion recognition systems become popularly used in everyday applications. For future growth, there are many questions that need to be addressed. This section briefly describes some of the challenges ahead in emotion recognition systems.

Data is of utmost importance. Having an appropriate database that is collected with a particular application and target user profile in mind can be expected to minimize the uncertainties and confusions that occur while organizing and labeling the database. Having high prototypical data clustered in well defined emotional spaces based on how they are perceived by target users will help to achieve optimal emotion recognition performance. The target-user-defined emotional spaces when combined with user profiles and

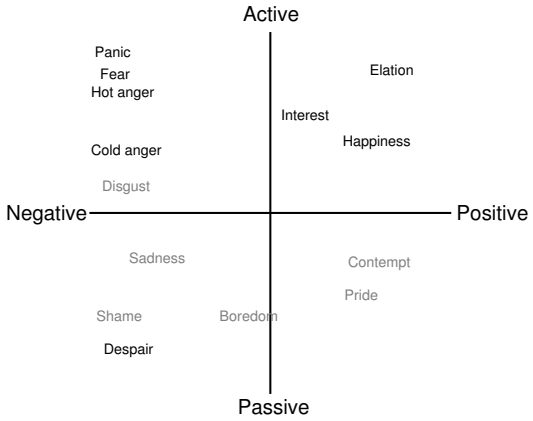


Fig. 7. Location of the emotional categories in the activation-valence space. For the emotional categories in gray, the power of the logistic regression model was inadequate to accurately recognize emotional from neutral speech ($r^2 < 0.5$). The figure was adapted from [23, 20].

data from other modalities (i.e., multi-modal emotional database consisting of many different sensor readings) will help to more effectively process real-life conditions and emotions.

As a first step, in view of the effort and cost required for data collection, existing general purpose spontaneous corpora can be better utilized. For example, huge corpora such as the Fisher English Training Speech corpus [18] and the Switchboard-I Telephone Speech Corpus [34] are likely to contain emotional content. With the help of automatic recognition systems, this content can be detected and studied to better understand spontaneous expression of emotions.

To assist and improve emotional evaluations, data and algorithms from different sources can be used to facilitate the process (human-in-the-loop). For example, Martin *et al.* proposed the use of image processing to annotate and validate emotional behaviors by quantizing the movement [48].

One area that should be further studied is the development, expression, perception, and progression of emotions in longer human-human or human-machine dialogs. The proposed framework should include explicit models of the context (i.e., emotions in previous turns, discourse information). Instead of modeling emotional category, the system could be designed to detect shifts in the emotional states of the users. If the application includes multi-person interaction, the framework should model the effect of the emotion of one user on the emotion states of the others.

Another area that should be studied is the design of adaptive emotion recognition systems. With the rapid development of mobile devices, it is expected that the demand for applications with emotional capabilities will in-

crease. In this context, the system should adapt to the specific emotional manifestations expressed by the user. Can we let the users or the applications choose the emotional labels? How can we easily modify and alter the labels to tailor them to specific applications and tools? How can we compensate for intercultural and inter-environment issues and differences? The answers to these questions are needed steps toward effective automatic speech emotion recognition systems.

There are many challenges and unknowns in research of recognizing emotions from speech. As in any research, it is essential to remember that even the smallest steps, which may seem unimportant, can be very important. For emotion recognition applications to flourish and become popular, we should design prototype systems to recognize emotions, even if they are only for constrained scenarios driven by concrete applications. In this direction, we proposed a real-time system with simple algorithms to extract and process spectral and prosodic acoustic features to detect negative speech [41]. Küstner proposed a demo emotional speech recognition system working in push-to-work mode [42]. There is also commercially available software for emotion identification named Layered Voice Analysis (LVA), which is being developed by Nemesysco Ltd. of Netanya. Only if this trend continues will we be able to explore the potential of human machine interfaces with emotional capabilities.

5 Summary

Emotions are the basic characteristics of humans and, therefore, incorporating them in applications, through recognition and synthesis, can improve the quality of life. In this chapter, we have described the characteristics of effective automatic speech emotion recognition systems. Specifically, database collection and organization; emotional descriptors; selection, calculation, and normalization of features; and training models were discussed to provide a summary of the current achievements, open questions, and future challenges.

Acknowledgment

This research was supported in part by funds from the NSF (through the Integrated Media Systems Center, a National Science Foundation Engineering Research Center, Cooperative Agreement No. EEC-9529152 and a CAREER award), the Department of the Army and a MURI award from ONR. Any opinions, findings and conclusions, or recommendations expressed in this book chapter are those of the authors and do not necessarily reflect the views of the funding agencies.

References

[1] S. Abrilian, L. Devillers, S. Buisine, and J.C.Martin. EmoTV1: Annotation of real-life emotions for the specification of multimodal affective interfaces. In *11th International Conference on Human-Computer Interaction (HCI 2005)*, pages 195–200, Las Vegas, Nevada, USA, July 2005.

[2] A. Alvarez, I. Cearreta, J.M. López, A. Arruti, E. Lazkano, B. Sierra, and N. Garay. Feature subset selection based on evolutionary algorithms for automatic emotion recognition in spoken Spanish and standard Basque language. In *Ninth International Conference on Text, Speech and Dialogue (TSD 2006)*, pages 565–572, Brno, Czech Republic, September 2006.

[3] R. Banse and K.R. Scherer. Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, 70(3):614–636, March 1996.

[4] A. Batliner, A. Buckow, H. Niemann, E. Nöth, and V. Warnke. The prosody module. In M.T. Maybury, O. Stock, and W. Wahlster, editors, *VERBMOBIL: Foundations of Speech-to-speech Translations*, pages 106–121. Springer Verlag, Berlin, Germany, 2000.

[5] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth. Desperately seeking emotions or: actors, wizards and human beings. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, pages 195–200, Newcastle, Northern Ireland, UK, September 2000.

[6] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth. How to find trouble in communication. *Speech Communication*, 40(1-2):117–143, April 2003.

[7] A. Batliner, S. Steidl, B. Schuller, D. Seppi, K. Laskowski, T. Vogt, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson. Combining efforts for improving automatic classification of emotional user states. In *Fifth Slovenian and First International Language Technologies Conference (IS-LTC 2006)*, pages 240–245, Ljubljana, Slovenia, October 2006.

[8] S. Biersack and V. Kempe. Tracing vocal emotion expression through the speech chain: do listeners perceive what speakers feel. In *ISCA Workshop on Plasticity in Speech Perception*, pages 211–214, London, UK, June 2005.

[9] F. Burkhardt, A. Paeschke, M. Rolfes, W.F. Sendlmeier, and B. Weiss. A database of German emotional speech. In *9th European Conference on Speech Communication and Technology (Interspeech'2005 - Eurospeech)*, pages 1517–1520, Lisbon, Portugal, September 2005.

[10] C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, and S.S. Narayanan. IEMOCAP: Interactive emotional dyadic motion capture database. *Journal of Language Resources and Evaluation*, 42(4):335–359, December 2008.

[11] C. Busso, M. Bulut, S. Lee, and S.S. Narayanan. Fundamental frequency analysis for speech emotion processing. In Sylvie Hancil, editor, *The Role*

of Prosody in Affective Speech, pages 309–337. Peter Lang Publishing Group, Berlin, Germany, 2009.

[12] C. Busso, S. Lee, and S.S. Narayanan. Using neutral speech models for emotional speech analysis. In *Interspeech 2007 - Eurospeech*, pages 2225–2228, Antwerp, Belgium, August 2007.

[13] C. Busso, S. Lee, and S.S. Narayanan. Analysis of emotionally salient aspects of fundamental frequency for emotion detection. *IEEE Transactions on Audio, Speech and Language Processing*, 17(4):582–596, May 2009.

[14] C. Busso and S.S. Narayanan. The expression and perception of emotions: Comparing assessments of self versus others. In *Interspeech 2008 - Eurospeech*, pages 257–260, Brisbane, Australia, September 2008.

[15] C. Busso and S.S. Narayanan. Recording audio-visual emotional databases from actors: a closer look. In *Second International Workshop on Emotion: Corpora for Research on Emotion and Affect, International conference on Language Resources and Evaluation (LREC 2008)*, pages 17–22, Marrakech, Morocco, May 2008.

[16] C. Busso and S.S. Narayanan. Scripted dialogs versus improvisation: Lessons learned about emotional elicitation techniques from the IEMOCAP database. In *Interspeech 2008 - Eurospeech*, pages 1670–1673, Brisbane, Australia, September 2008.

[17] J.P. Campbell. Speaker recognition: a tutorial. *Proceedings of the IEEE*, 85(9):1437–1462, September 1997.

[18] C. Cieri, D. Miller, and K. Walker. The Fisher corpus: A resource for the next generations of speech-to-text. In *International conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal, May 2004.

[19] C. Clavel, I. Vasilescu, L. Devillers, G. Richard, and T. Ehrette. Fear-type emotion recognition for future audio-based surveillance systems. *Speech Communication*, 50(6):487–503, June 2008.

[20] R. Cowie and R.R. Cornelius. Describing the emotional states that are expressed in speech. *Speech Communication*, 40(1-2):5–32, April 2003.

[21] R. Cowie, E. Douglas-Cowie, and C. Cox. Beyond emotion archetypes: Databases for emotion modelling using neural networks. *Neural Networks*, 18(4):371–388, May 2005.

[22] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder. 'FEELTRACE': An instrument for recording perceived emotion in real time. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, pages 19–24, Newcastle, Northern Ireland, UK, September 2000. ISCA.

[23] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J.G. Taylor. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 18(1):32–80, January 2001.

[24] J.R. Deller, J.H.L. Hansen, and J.G. Proakis. *Discrete-Time Processing of Speech Signals*. IEEE Press, Piscataway, NJ, USA, 2000.

[25] L. Devillers and L. Vidrascu. Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs. In *Interspeech - International Conference on Spoken Language (ICSLP)*, pages 801–804, Pittsburgh, PA, USA, September 2006.

[26] L. Devillers, L. Vidrascu, and L. Lamel. Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks*, 18(4):407–422, May 2005.

[27] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach. Emotional speech: Towards a new generation of databases. *Speech Communication*, 40(1-2):33–60, April 2003.

[28] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, J.-C. Martin, L. Devillers, S. Abrilian, A. Batliner, N. Amir, and K. Karpouzis. The HUMAINE database: Addressing the collection and annotation of naturalistic and induced emotional data. In A. Paiva, R. Prada, and R.W. Picard, editors, *Affective Computing and Intelligent Interaction*, pages 488–500. Springer-Verlag Press, Berlin, Germany, September 2007.

[29] E. Douglas-Cowie, L. Devillers, J.C. Martin, R. Cowie, S. Savvidou, S. Abrilian, and C. Cox. Multimodal databases of everyday emotion: Facing up to complexity. In *9th European Conference on Speech Communication and Technology (Interspeech'2005)*, pages 813–816, Lisbon, Portugal, September 2005.

[30] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification*. Wiley-Interscience, New York, NY, USA, 2000.

[31] F. Enos and J. Hirschberg. A framework for eliciting emotional speech: Capitalizing on the actors process. In *First International Workshop on Emotion: Corpora for Research on Emotion and Affect (International conference on Language Resources and Evaluation (LREC 2006))*, pages 6–10, Genoa,Italy, May 2006.

[32] D. Foley. Considerations of sample and feature size. *IEEE Transactions on Information Theory*, 18(5):618–626, September 1972.

[33] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dahlgren. Timit acoustic-phonetic continuous speech corpus, 1993.

[34] J.J. Godfrey and E. Holliman. Switchboard-1 release 2, 1997. Linguistic Data Consortium.

[35] J. Gratch, S. Marsella, and P. Petta. Modeling the cognitive antecedents and consequences of emotion. *Journal of Cognitive Systems Research*, 10(1):1–5, March 2009.

[36] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan. Primitives-based evaluation and estimation of emotions in speech. *Speech Communication*, 49(10-11):787–800, October-November 2007.

[37] M. Grimm, K. Kroschel, and S. Narayanan. The Vera AM Mittag German audio-visual emotional speech database. In *IEEE International Conference on Multimedia and Expo (ICME 2008)*, pages 865–868, Hannover, Germany, June 2008.

[38] Humaine project portal. http://emotion-research.net/, 2009. Retrieved March 31st, 2009.

[39] O. Kalinli and S. Narayanan. Early auditory processing inspired features for robust automatic speech recognition. In *XV European Signal Processing Conference (EUSIPCO 2007)*, pages 2385–2389, Poznan, Poland, September 2007.

[40] A. Kazemzadeh, S. Lee, and S. Narayanan. An interval type-2 fuzzy logic system to translate between emotion-related vocabularies. In *Interspeech 2008 - Eurospeech*, pages 2747–2750, Brisbane, Australia, September 2008.

[41] S. Kim, P.G. Georgiou, S. Lee, and S.S. Narayanan. Real-time emotion detection system using speech:multi-modal fusion of different timescale features. In *International Workshop on Multimedia Signal Processing (MMSP 2007)*, pages 48–51, Chania, Crete, Greece, October 2007.

[42] O. Küstner, R. Tato, T. Kemp, and B. Meffert. Towards real life applications in emotion recognition. In E. André, L. Dybkaer, W. Minker, and P. Heisterkamp, editors, *Affective Dialogue Systems(ADS 2005), Lecture Notes in Artificial Intelligence 3068*, pages 25–35. Springer-Verlag Press, Berlin, Germany, May 2004.

[43] C.M. Lee and Shrikanth Narayanan. Emotion recognition using a data-driven fuzzy inference system. In *8th European Conference on Speech Communication and Technology (Eurospeech 2003)*, pages 157–160, Geneva, Switzerland, September 2003.

[44] C.M. Lee and S.S. Narayanan. Toward detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing*, 13(2):293–303, March 2005.

[45] C.M. Lee, S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S.S. Narayanan. Emotion recognition based on phoneme classes. In *8th International Conference on Spoken Language Processing (ICSLP 04)*, pages 889–892, Jeju Island, Korea, October 2004.

[46] S. Lee, S. Yildirim, A. Kazemzadeh, and S. Narayanan. An articulatory study of emotional speech production. In *9th European Conference on Speech Communication and Technology (Interspeech'2005 - Eurospeech)*, pages 497–500, Lisbon, Portugal, September 2005.

[47] S.C. Marsella and J. Gratch. EMA: A process model of appraisal dynamics. *Journal of Cognitive Systems Research*, 10(1):70–90, March 2009.

[48] J.-C. Martin, G. Caridakis, L. Devillers, K. Karpouzis, and S. Abrilian. Manual annotation and automatic image processing of multimodal emotional behaviors: validating the annotation of TV interviews. *Personal and Ubiquitous Computing*, 13(1):69–76, January 2009.

[49] Mirex 2009. http://www.music-ir.org/mirex/2009/index.php, 2009. Retrieved March 31st, 2009.

[50] I.R. Murray and J.L. Arnott. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *Journal of the Acoustical Society of America*, 93(2):1097–1108, February 1993.

[51] National Institute of Standards and Technology, spoken language technology evaluations. www.nist.gov/speech/tests/, 2009. Retrieved March 31st, 2009.

[52] M. Pantic and L.J.M. Rothkrantz. Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE*, 91(9):1370–1390, September 2003.

[53] T.L. Pao, W.Y. Liao, Y.T. Chen, J.H. Yeh, Y.M. Cheng, and C.S. Chien. Comparison of several classifiers for emotion recognition from noisy Mandarin speech. In *Third International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIHMSP 2007)*, volume 1, pages 23–26, Kaohsiung, Taiwan, November 2007.

[54] D.B. Paul and J.M. Baker. The design for the Wall Street Journal-based CSR corpus. In *2th International Conference on Spoken Language Processing (ICSLP 1992)*, pages 899–902, Banff, Alberta, Canada, October 1992.

[55] R. W. Picard. Affective computing. Technical Report 321, MIT Media Laboratory Perceptual Computing Section, Cambridge, MA,USA, November 1995.

[56] K.R. Scherer. Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40(1-2):227–256, April 2003.

[57] K.R. Scherer, R. Banse, and H.G. Wallbott. Emotion inferences from vocal expression correlate across languages and cultures. *Journal of Cross-Cultural Psychology*, 32(1):76, January 2001.

[58] K.R. Scherer and G. Ceschi. Lost luggage: A field study of emotion antecedent appraisal. *Motivation and Emotion*, 21(3):211–235, September 1997.

[59] F. Schiel, S. Steininger, and U. Türk. The SmartKom multimodal corpus at BAS. In *Language Resources and Evaluation (LREC 2002)*, Las Palmas, Spain, May 2002.

[60] B. Schuller, D. Arsic, F. Wallhoff, and G. Rigoll. Emotion recognition in the noise applying large acoustic feature sets. In *ISCA Speech Prosody*, Dresden, Germany, May 2006. ISCA.

[61] B. Schuller, A. Batliner, D. Seppi, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson. The relevance of feature type for the automatic classification of emotional user states: Low level descriptors and functionals. In *Interspeech 2007 - Eurospeech*, pages 2253–2256, Antwerp, Belgium, August 2007.

[62] B. Schuller, R. Müeller, B. Höernler, A. Höethker, H. Konosu, and G. Rigoll. Audiovisual recognition of spontaneous interest within conversations. In *9th international conference on Multimodal interfaces (ICMI 2007)*, pages 30–37, Nagoya, Aichi, Japan, November 2007.

[63] B. Schuller, D. Seppi, A. Batliner, A. Maier, and S. Steidl. Towards more reality in the recognition of emotional speech. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2007)*, volume 4, pages 941–944, Honolulu, HI, USA, April 2007.

[64] B. Schuller, S. Steidl, and A. Batliner. The INTERSPEECH 2009 emotion challenge. In *Interspeech 2009 - Eurospeech*, Brighton, UK, September 2009.

[65] M.H. Sedaaghi, C. Kotropoulos, and D. Ververidis. Using adaptive genetic algorithms to improve speech emotion recognition. In *International Workshop on Multimedia Signal Processing (MMSP 2007)*, pages 461–464, Chania, Crete, Greece, October 2007.

[66] D. Seppi, A. Batliner, B. Schuller, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, and V. Aharonson. Patterns, prototypes, performance: Classifying emotional user states. In *Interspeech 2008 - Eurospeech*, pages 601–604, Brisbane, Australia, September 2008.

[67] S. Steidl. *Automatic Classification of Emotion-Related User States in Spontaneous Childrens Speech.* PhD thesis, Universität Erlangen-Nürnberg, Erlangen, Germany, January 2009.

[68] S. Steidl, M. Levit, A. Batliner, E. Nöth, and H. Niemann. "Of all things the measure is man" automatic classification of emotions and inter-labeler consistency. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005)*, volume 1, pages 317–320, Philadelphia, PA, USA, March 2005.

[69] K.P. Truong, M.A. Neerincx, and D.A. van Leeuwen. Assessing agreement of observer- and self-annotations in spontaneous multimodal emotion data. In *Interspeech 2008 - Eurospeech*, pages 318–321, Brisbane, Australia, September 2008.

[70] D. Ververidis and C. Kotropoulos. Fast sequential floating forward selection applied to emotional speech features estimated on DES and SUSAS data collections. In *XIV European Signal Processing Conference (EUSIPCO 2006)*, pages 929–932, Florence, Italy, September 2006.

[71] L. Vidrascu and L. Devillers. Real-life emotions in naturalistic data recorded in a medical call center. In *First International Workshop on Emotion: Corpora for Research on Emotion and Affect (International conference on Language Resources and Evaluation (LREC 2006))*, pages 20–24, Genoa,Italy, May 2006.

[72] S. Yildirim, M. Bulut, C.M. Lee, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S.S. Narayanan. An acoustic study of emotions expressed in speech. In *8th International Conference on Spoken Language Processing (ICSLP 04)*, pages 2193–2196, Jeju Island, Korea, October 2004.

[73] Z. Zeng, M. Pantic, G.I. Roisman, and T.S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58, January 2009.