Oxford Handbooks Online

Speech in Affective Computing

Chi-Chun Lee, Jangwon Kim, Angeliki Metallinou, Carlos Busso, Sungbok Lee, and Shrikanth S. Narayanan

The Oxford Handbook of Affective Computing
Edited by Rafael Calvo, Sidney D'Mello, Jonathan Gratch, and Arvid Kappas

Print Publication Date: Jan 2015 Subject: Psychology, Affective Science
Online Publication Date: Jul 2014 DOI: 10.1093/oxfordhb/9780199942237.013.021

Abstract and Keywords

This chapter is from the forthcoming *The Oxford Handbook of Affective Computing* edited by Rafael Calvo, Sidney K. D'Mello, Jonathan Gratch, and Arvid Kappas. Speech is a key communication modality for humans to encode emotion. In this chapter, we address three main aspects of speech in affective computing: emotional speech production, acoustic feature extraction for emotion analysis, and the design of a speech-based emotion recognizer. Specifically we discuss the current understanding of the interplay of speech production vocal organs during expressive speech, extracting informative acoustic features from speech recording waveforms, and the engineering design of automatic emotion recognizers using speech acoustic-based features. The latter includes a discussion of emotion labeling for generating ground truth references, acoustic feature normalization for controlling signal variability, and choice of computational frameworks for emotion recognition. Finally, we present some open challenges and applications of a robust emotion recognizer.

Keywords: emotional speech production, acoustic feature extraction for emotion analysis, computational frameworks for emotion recognition, acoustic feature normalization

Introduction

Speech is a natural and rich communication medium for humans to interact with one another. It encodes both linguistic intent and paralinguistic information (e.g., emotion, age, gender, etc.). In this chapter, we focus our discussion on this unique human behavior modality, speech, in the context of affective computing, in order to measure and quantify the internal emotional state of a person by observing external affective and expressive

behaviors. The specific focus is on describing the emotional encoding process in speech production—that is,the state-of-the-art computational approaches and future directions and applications of computing affect from speech signals.

The human speech signal is a result of complex and integrative movement of various speech production organs including the vocal chords, larynx, pharynx, tongue, velum, and jaw. With the availability of instrumental technologies—including ultrasound, x-ray microbeam, electromagnetic articulography (EMA), and (real time) magnetic resonance imaging (MRI)—researchers have begun to investigate various scientific questions in order to bring insights into the emotional speech production mechanisms. In this chapter, we start by providing some empirical details of how emotional information is encoded at the speech production level (Affective Speech Production, p. 171).

Research in understanding the production mechanisms of emotional speech is still evolving. However, empirical computational approaches for extracting acoustic signal features that characterize emotional speech have emerged from scientific advances both in emotion perception and speech signal analysis. In Computation of Affective Speech Features (p. 173), we summarize the set of features of vocal cues that have become a de facto standard, often termed as the speech low-level descriptors (LLDs), for automatic emotion recognition.

(p. 171) In Affect Recognition and Modeling Using Speech (p. 175), we describe three essential components of a proper design of an automatic emotion recognition system using speech-acoustic features: definition and implementation of emotion labeling that serve as the basis for computing (Emotion Labels for Computing, p. 175), acoustic feature normalization that helps address issues related to signal variability due to factors other than the core emotions being targeted (Robust Acoustic Feature Normalization, p. 176), and machine learning algorithms that offer the means for achieving the desired modeling goal (Computational Framework for Emotion Recognition, p. 177).

Emotion labeling (or annotation) typically provides a ground truth for training and evaluating emotion recognition systems. The specific choice of representations (descriptors) used for computing depends on the theoretical underpinnings and the application goal. In addition to traditionally used categorical (happy, angry, sad, and neutral) and dimensional labels (of arousal, valence, and dominance), researchers have made advances in computationally integrating behavior descriptors in the characterization of emotion. These advancements can better handle the ambiguity in the definition of emotions compared with traditional labeling schemes (Emotion Labels for Computing, p. 175).

Normalization of acoustic features aims to minimize unwanted variability due to sources other than the construct (i.e., emotion) being modeled. The speech signal is influenced by numerous factors including what is being said (linguistic content), who is saying it (speaker identity, age, gender), how the signal is being captured and transmitted (telephone, cellphone, microphone types), and the context in which the speech signal is generated (room acoustics and environment effects including background noise). In

Robust Acoustic Feature Normalization, p. 176, we discuss several techniques for feature normalization that ensure that the features contain more information about emotion and less about other nonemotional confounding variability.

Machine learning algorithms are used to train the recognition system to learn a mapping between the extracted speech features and the given target emotion labels. Many standard pattern recognition techniques used in other engineering applications have shown to be appropriate for emotion recognition system with speech features. We also describe other recent state-of-the-art emotion recognition frameworks that have been proposed to take into account of the various contextual influences in the expression of emotions in speech, including the nature of human interactions for obtaining improved emotion recognition accuracies (Computational Framework for Emotion Recognition, p. 177).

There remain many challenges that require further investigation and future research; however, potential engineering applications, including new generation of human-machine interfaces, have made the development of robust emotion-sensing technology essential. A recent research endeavor of the rapidly growing field of behavior signal processing (BSP) (Narayanan & Georgiou, 2013) has demonstrated that development can provide analytical tools for advancing behavioral analyses desired by domain experts across a wide range of disciplines, especially in fields related to mental health (Speech in Affective Computing: Future Works and Applications, p. 180).

Affective Speech Production

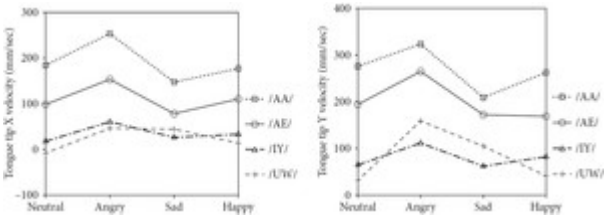
Often, speech production research is conducted under the “source filter” theory (Fant, 1970), which views the speech production system as consisting of two components: source activities, which generate airflow, and vocal tract shaping filtering, which modulates the airflow. Although laryngeal behavior is not fully independent of supralaryngeal elements, the modulation of vocal folds, or vocal cords, in the larynx is the primary control of source activity. This modulation results in the variation of pitch (the frequency of vocal fold vibration), intensity (the pressure of the airflow), and voice quality dynamics (degrees of aperiodicity in the resulting glottal cycle). Note that the filter affects the variation of intensity and voice quality, too. The air stream passed through the vocal fold is modulated by articulatory controls of tongue, velum, lips, and jaw in the vocal tract, resulting in dynamic spectral changes in the speech signal. The interaction and interplay between voice source activities and articulatory controls also contribute to the speech sound modulation.

Most emotional speech studies have focused on the acoustic characteristics of the resulting speech signal level—such as the underlying prosodic variation, spectral shape, and voice quality change—across various time scales rather than considering the underlying production mechanisms directly. In order to understand complex acoustic

structure and further the human communication process that involves information encoding and decoding, a deeper understanding of orchestrated articulatory activity is needed. In this section, we describe scientific findings of emotional speech production in terms of articulatory mechanisms, vocal folds actions, and the interplay between voice source and articulatory kinematics.

(p. 172) Articulatory Mechanisms in Emotionally Expressive Speech

The number of studies on articulatory mechanisms of expressive speech is limited compared with studies in the acoustic domain presumably due to the difficulties in obtaining direct articulatory data. Contemporary instrumental methods for collecting articulatory data include ultrasound (Stone, 2005), x-ray microbeam (Fujimura, Kiritani, & Ishida, 1973), electromagnetic articulography (EMA) (Perkell et al., 1992), and (real time) magnetic resonance imaging (MRI) (Narayanan, Alwan, & Haker, 1995; Narayanan, Nayak, Lee, Sethy, & Byrd, 2004). While it is often challenging for subjects to express emotions naturally in these data collection environments, there have been some systematic studies with these data collection technologies showing that articulatory patterns of acted emotional speech are different from neutral (nonemotional) speech.



Click to view larger

Fig. 12.1 Tongue tip horizontal (left) and vertical (right) movement velocity plots of four peripheral vowels as a function of emotion.

Source: Lee, Yildirim, Kazemzadeh, and Narayanan (2005).

Lee et al. analyzed the surface articulatory motions by using emotional speech data for four acted emotions (angry, happy, sad, and neutral) collected with EMA (Lee, Yildirim, Kazemzadeh, & Narayanan, 2005). The study showed that that the speech production of emotional speech is associated more with peripheral articulatory motions than that of neutral speech. For example, the tongue tip (TT), jaw, and lip positioning are more advanced (extreme) in emotional speech than in neutral speech (Figure 12.1). Furthermore, the results of multiple simple discriminant analyses treating the four emotion categories as dependent variable showed that the classification recalls of using articulatory features are higher than those of acoustic features. The result implied that the articulatory features carry valuable emotion-dependent information.

Lee et al. also found that there was more prominent usage of the pharyngeal region for anger than neutral, sadness and happiness in emotional speech (Lee, Bresch, Adams, Kazemzadeh, & Narayanan, 2006). It was further observed that happiness is associated with greater laryngeal elevation than anger, neutrality, and sadness. This emotional variation of the larynx was related to wider pitch and second formant (F2) ranges and

higher third formant frequencies (F3) in the acoustic signal. It was also reported that the variation of articulatory positions and speed as well as pitch and energy are significantly associated with perceptual strength of emotion in general (Kim, Lee, & Narayanan, 2011).

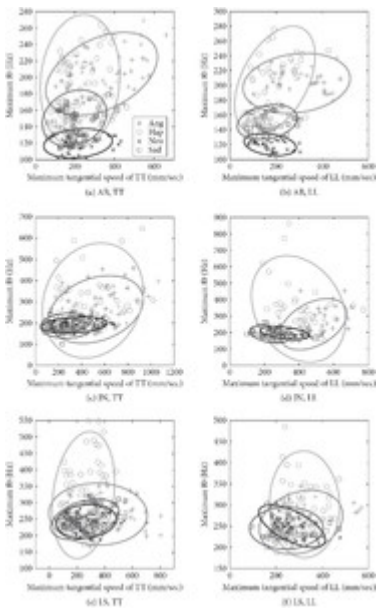
Most of the emotional speech production studies rely on acted emotion recorded using actors/actresses as subjects. Although acted emotional speech could be different from spontaneous emotional speech in terms of articulatory positions (Erickson, Menezes, & Fujino, 2004), using acted emotional expression remains one of the most effective methods for collecting articulatory data in order to carry out studies in emotional speech production. A certain degree of ecological validity is achieved by following consistent experimental techniques such as those expressed by Busso and Narayanan (Busso & Narayanan, 2008).

Vocal Fold Controls in Emotionally Expressive Speech

Vocal fold controls or, more precisely, the controls of the tension and length of vocal fold muscles, enable major modulations of voice source activities. Voice source is defined as the airflow passing through the glottis in the larynx. The configuration of voice source is determined by the actions of opening and closing of vocal folds (p. 173) with different levels of tensions in the laryngeal muscles. During speech production, the voice source is filtered by supralaryngeal vocal organs. Since the speech waveform is the result of complex modulations (filtering) of glottal airflow in the supraglottal structure, it is difficult to recover the glottal airflow information from the speech output acoustics. One of the most popular techniques to recover voice source is through inverse filtering; however, it remains challenging to estimate the voice source information from natural spontaneous speech even with little noise and distortion.

Despite these difficulties, there are interesting studies reporting on paralinguistic aspects of voice source activities in the domain of emotional speech production. For example, for sustained /aa/, Murphy et al. showed that the estimated contacting quotient (i.e., contact time of the vocal folds divided by cycle duration) and speed quotient, or velocity of closure divided by velocity of opening, from the electroglottogram (EGG), are different among five categorical (simulated) emotions (angry, joy, neutrality, sadness, and tenderness) (Murphy & Laukkanen, 2009). Gobl et al. also showed that voice qualities—such as harsh, tense, modal, breathy, whispery, creaky and lax-creaky, and combinations of them—are associated with affective states using synthesized speech (Gobl & Chasaide, 2003).

Interplay Between Voice Source and Articulatory Kinematics



Click to view larger

Fig. 12. 2 Example plots of the maximum tangential speed of critical articulators and the maximum pitch. A circle indicates that Gaussian contour with 2 sigma standard deviation for each emotion (red-Ang, green-Hap, black-Neu, blue-Sad). Different emotions show distinctive variation patterns in the articulatory speed dimension and the pitch dimension.

Source: Kim, Lee, and Narayanan (2010).

Another essential source of emotional information in speech production is present in the interplay between voice source activities and articulatory kinematics. Kim et al. reported that angry speech introduces the greatest articulatory speed modulations, while pitch modulations were most prominent for happy speech (Kim, Lee, & Narayanan, 2010) (Figure 12.2). This study underscores the complexity and the importance of better understanding the interplay between voice source behavior and articulatory motion in the analysis of emotional speech production.

Open Challenges

One of the biggest challenges and opportunities in studying emotional variation in speech production lies in the inter- and intraspeaker variability. Interspeaker variability includes heterogeneous display of emotion and differences in individual's vocal tract structures (Lammert, Proctor, & Narayanan, 2013). Intraspeaker variability results from the fact that a speaker can express an emotion in a number of ways and is influenced by the context. The invariant nature of controls of speech production components still remains elusive, making comprehensive modeling of emotional speech challenging and largely open.

Computation of Affective Speech Features

As described in Affective Speech Production (p. 171), the analysis of speech production data suggests that a complex interaction between vocal source activities and vocal tract modulations likely underlies how emotional information is encoded in speech waveform.

While an understanding of this complex emotional speech production mechanism is emerging only as more research is being carried out, many studies have examined the relationship between the perceptual quality of emotional content and acoustic signal characteristics.

Bachorowski has summarized a wide range of results from various psychological perceptual tests indicating that humans are significantly more accurate at judging emotional content than merely guessing at chance level while listening to speech recordings (Bachorowski, 1999). Furthermore, Scherer described a comprehensive theoretical production-perception model of vocal communication of emotion and provided a detailed review on how each acoustic parameter (e.g., pitch, intensity, speech rate, etc.) covaries with different intensities of emotion perception (Scherer, 2003); this classic study was further expanded upon in the handbook for nonverbal behavior research focusing on the vocal expression of affect (Juslin & Scherer, 2005). These studies of the processing of emotional speech by humans have formed the bases for affective computing using speech owing to its extensive scientific grounding. They have also served as an initial foundation for developing engineering applications of affective computing (e.g., emotion recognition using speech and emotional speech synthesis).

Acoustic Feature Extraction for Emotion Recognition

Computing affect from speech signals has benefited greatly from the perceptual understanding and, to a smaller extent, the production details of vocal expressions and affect. A list of commonly used acoustic low-level descriptors (LLDs), extracted from speech recordings that can be used in emotion recognition tasks is given below. (p. 174)

(p. 175) Prosody-related signal measures

- Fundamental frequency (f_0)
- Short-term energy
- Speech rate: syllable/phoneme rate

Spectral characteristics measures

- Mel-frequency cepstral coefficients (MFCCs)
- Mel-filter bank energy coefficients (MFBs)

Voice quality–related measures

- Jitter
- Shimmer
- Harmonic-to-noise ratio

Prosody relates to characteristics such as rhythm, stress, and intonation of speech; spectral characteristics are related to the harmonic/resonant structures resulting as the airflow is modulated by dynamic vocal tract configurations; and voice quality measures are related to the characteristics of vocal fold vibrations (e.g., degrees of aperiodicity in the resulting speech waveform).

Many publicly available toolboxes are capable of performing such acoustic feature computation. OpenSmile (Eyben, Wöllmer, & Schuller, 2010) is one such toolbox designed specifically for emotion recognition tasks; other generic audio/speech processing toolboxes—such as Praat (Boersma, 2001), Wavesurfer,[1] and Voicebox[2]—are all capable of extracting relevant acoustic features.

In practice, after extracting these LLDs, researchers frequently further apply a data processing approach, often computed at a time-scale of 10 to 25 milliseconds, in order to capture the rich dynamics. The approach first involves computing various statistical functionals (i.e., mean, standard deviation, range, interquartile range, regression residuals, etc.) on these LLDs at different time scale granularities (e.g., at 0.1, 0.5, 1, and 10 seconds, etc.). Furthermore, in order to measure the dynamics at multilevel time scales, statistical functional operators can also be stacked on top of each other; for example, one can compute the mean of pitch LLDs (i.e., fundamental frequency) for every 0.1 second, then compute the mean of "the mean of pitch (at 0.1s)" for every 0.5 second, and repeat this process with increasing time scales across different statistical functional operators.

This data processing technique has been applied successfully in tasks such as emotion recognition (Lee, Mower, Busso, Lee, & Narayanan, 2011; Schuller, Arsic, Wallhoff, & Rigoll, 2006; Schuller, Batliner, et al., 2007), paralinguistic prediction (Bone, Li, Black, & Narayanan, 2012; Björn Schuller et al., 2013), and other behavioral modeling (Black et al., 2013; Black, Georgiou, Katsamanis, Baucom, & Narayanan, 2011). This approach can often result in a very high-dimensional feature vector—for example, depending on the length of audio segment, it can range from hundreds of features to thousands or more. Feature selection techniques—stand-alone (e.g., correlation-based) (Hall, 1999) and mutual information based (Peng, Long, & Ding, 2005) or wrapper selection techniques (e.g., sequential forward feature selection, sequential floating forward feature selection) (Jain & Zongker, 1997)—can be carried out to reduce the dimension appropriately for the set of emotion classes of interest.

Open Challenges

While the aforementioned data processing approach has been shown to be effective in various emotion prediction tasks, it remains unclear why the large number of acoustic LLDs work well and what aspects of emotional production-perception mechanisms are captured with this technique. From a computational point of view, since it is an exhaustive and computationally expensive approach, an efficient and reliable real-life

emotional recognizer built upon this approach may be impractical. Future works lie in designing better-informed features based on the understanding of emotional speech production-perception mechanisms while maintaining reliable prediction accuracies compared with the current approach.

Affect Recognition and Modeling Using Speech

Recognizing and tracking emotional states in human interactions based on spoken utterances requires a series of appropriate engineering design including the following: specifying an annotation scheme of appropriate emotion labels, implementing a feature normalization technique for robust recognition, and designing context-aware machine learning frameworks to model the temporal and interaction aspect of emotion evolution in dialogs.

Emotion Labels for Computing

Annotating (coding) data with appropriate emotion labels is a crucial first step in providing the basis for implementing and evaluating the computational modeling approaches. Traditionally, behavioral assessment of one's emotional state (p. 176) can be done in two different ways: self-reports or perceived ratings. Self-reported emotion assessment instruments are designed to ask the subjects to recall his or her experience and memory about how he or she has felt during a particular interaction (e.g., the positive and negative affect schedule (PANAS) (Watson, Clark, & Tellegen, 1988). Perceived-ratings are often carried out by asking external (trained) observers to assign labels of emotion as they watch a given audiovideo recording. Tools such as ELAN (Wittenburg, Brugman, Russel, Klassmann, & Sloetjes, 2006) and Anvil (Kipp, 2001) are commonly used software for carrying out such annotations.

Many studies of emotion in behavioral science rely on self-assessment of emotional states to approximate the true underlying emotional states of the subject. This method of emotion labeling is often used to clarify the role of human affective process under different scientific hypotheses. In affective computing, recognizing emotion automatically from recorded behavioral data often adopts annotation based on perceived emotion. The perceived emotional states can be coded either as categorical emotional states (e.g., angry, happy, sad, neutral) or as dimensional representations (e.g., valence, activation, and dominance). This method of labeling emotion is motivated by the premise that automatic emotion recognition systems are often designed with an aim of recognizing emotions through perceiving/sensing other humans' behaviors.

Depending on the applications, one can take an approach of labeling behavioral data with self-reported assessment instrument or perceived emotional states. The design of labeling serves as ground truth for training and testing machine learning algorithms and the

choice of different labeling schemes also often comes with a distinct interpretation of whether the model is capturing the underlying human affective production or perception process.

Recent Advances in Emotion Labeling

Many of the traditional emotion labels can be seen as a compact representation of a large emotion space. Individual differences in internalizing what constitutes a specific emotion label often arise from the variation of an integrative process of cognitive evaluation of personal experience and spontaneous behavioral reaction to affective stimuli. There are some recent computational works aimed at advancing representations of emotions by incorporating signal-based behavior descriptors that are more conducive to capture the nonprototypical blended nature in real life (Mower et al., 2009). A recent work demonstrated the representation of emotion as emotion profile (i.e., a mixture of categorical emotional labels based on models built with visual-acoustic descriptors). This approach can model the inherent ambiguity and subtle nature of emotional expressions (Mower, Mataric, & Narayanan, 2011). Another recent representation in exploring computational method to better represent this large emotion space is through the use of natural language (Kazamzadeh, Lee, Georgiou, & Narayanan, 2011). This approach aims at representing any emotion word in terms of humans' natural language either describing a past event, a memorable experience, or simply closely related traditionally used categorical emotional states.

Robust Acoustic Feature Normalization

Speech is a rich communication medium conveying emotional, lexical, cultural, and idiosyncratic information, among others, and it is often affected by the environment (e.g., noise, reverberation) and recording and signal transmission setup (e.g., microphone quality, sampling rate, wireless/VoIP channels, etc.).

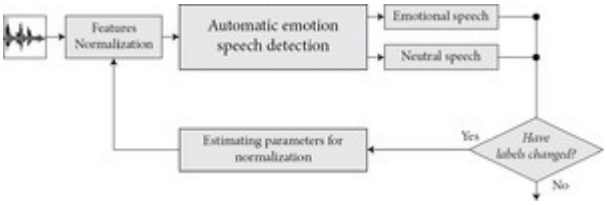
Previous studies have indicated the importance of speaker normalization in recognizing paralinguistic information (Bone, Li, et al., 2012; Busso, Lee, & Narayanan, 2009; Rahman & Busso, 2012). For example, the structure and the size of the larynx and the vocal folds determine the values of the fundamental frequency (f_0), which span the range of 50 to 250Hz for men, 120 to 500Hz for women, and even higher for children (Deller, Hansen, & Proakis, 2000). Therefore, although angry speech has a higher f_0 values than neutral speech (Yildirim et al., 2004), the emotional differences can be blurred by interspeaker differences—the difference between the mean values of the fundamental frequency of neutral and anger speech during spontaneous interaction (e.g., the USC IEMOCAP database (Busso et al., 2008) is merely a 68-Hz shift.

A common approach to normalize the data is to estimate global acoustic parameters across speakers and utterances. For example, the z-normalization approach transforms the features by subtracting their mean and dividing by their standard deviation (i.e., each feature will have zero mean and unit variance across all data) (Lee & Narayanan, (p. 177)

2005; Lee et al., 2011; Metallinou, Katsamanis, & Narayanan, 2012; Schuller, Rigoll, & Lang, 2003). The min-max approach scales the feature to a predefined range (Clavel, Vasilescu, Devillers, Richard, & Ehrette, 2008; Pao, Yeh, Chen, Cheng, & Lin, 2007; Wöllmer et al., 2008). Other nonlinear normalization approaches aim to convert the features' distributions into normal distributions (Yan, Li, Cairong, & Yinhua, 2008). Studies have applied these approaches in speaker-dependent conditions in which the normalization parameters are separately estimated for each individual (Bitouk, Verma, & Nenkova, 2010; Le, Quénot, & Castelli, 2004; Schuller, Vlasenko, Minguez, Rigoll, & Wendemuth, 2007; Sethu, Ambikairajah, & Epps, 2007; Vlasenko, Schuller, Wendemuth, & Rigoll, 2007; Wöllmer et al., 2008).

Iterative Feature Normalization (IFN)

Busso et al. demonstrated that global normalization is not always effective in increasing the performance of an emotion recognition system (Busso, Metallinou, & Narayanan, 2011). This is because applying a single normalization scheme across the entire corpus can adversely affect the emotional discrimination of the features (e.g., all features having the same mean and range across sentences). A new transformation is done by normalizing features by estimating the parameters of an affine transformation (e.g., z-normalization) using only neutral (nonemotional) samples.



Click to view larger

Fig. 12.3 Iterative feature normalization. This unsupervised front end uses an automatic emotional speech detector to identify neutral samples, which are used to estimate the normalization parameters. The process is iteratively repeated until the labels are not modified.

Source: Busso, Metallinou, and Narayanan (2011).

Multiple studies have consistently observed statistically significant improvements in performance (Busso et al., 2009, 2011; Rahman & Busso, 2012) when this approach is separately applied for each subject. Given that neutral samples may not be available for each of the target individual, Busso et al. proposed the iterative feature normalization (IFN) scheme (Figure 12.3) (Busso et al., 2011). This unsupervised front-end scheme implements the aforementioned ideas by estimating the neutral subset of the data iteratively and using this partition to estimate the normalization parameters. As the features are better normalized, the emotion detection system provides more reliable estimation, which, in turn, produces better normalization parameters. The IFN approach is also robust against different recording conditions, achieving over 19% improvement in unweighted accuracy (Rahman & Busso, 2012).

Computational Framework for Emotion Recognition

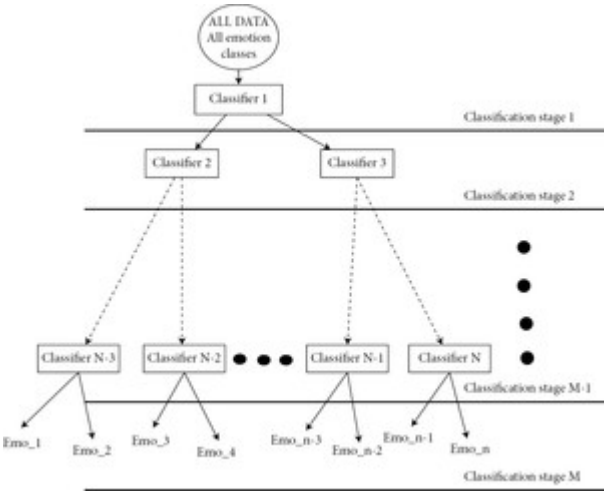
Supervised machine learning algorithms are at the heart of many emotion recognition efforts. These machine learning algorithms map input behavioral descriptions (automatically derived acoustic features, Acoustic Feature Extraction for Emotion Recognition, p. 173) through normalization (Robust Acoustic Feature Normalization, p. 176) to desired emotion representations (emotional labeling, Emotion Labels for Computing, p. 175).

An excellent survey of the various machine learning methodologies of affective modeling can be found in Zeng, Pantic, Roisman, and Huang (2009). If an input signal is given an emotion label using categorical attributes, many state-of-the-art static classifiers (e.g., support vector machine, decision tree, naive Bayes, hidden Markov model, etc.) can be implemented directly as the basic classifier. Furthermore, when an utterance is evaluated based on dimensional representation (i.e., valence, activation, and dominance), various well-established regression techniques such as ordinary/robust least square regression and support vector regression, can be utilized. Publicly available machine learning toolboxes such as WEKA (Hall et al., 2009), LIBSVM (Chang & Lin, 2011), and HTK (Young et al., (p. 178) 2006) have implemented the above-mentioned classification/ regression techniques and are widely used.

In this section, we discuss three different exemplary, recently developed novel emotion recognition frameworks for automatically recognizing emotional attributes from speech: The first is a *static* emotion attributes classification system based on a binary decision hierarchical tree structure, the second comprises two *context*-sensitive frameworks for emotion recognition in dialogues, and the third is a framework for continuous evaluation of emotion flow in human interactions.

Static Emotion Recognition for Single Utterance

In order to map an individual input utterance to a predefined set of categorical emotion classes given acoustic features, an exemplary approach is a hierarchical tree-based approach (Lee et al., 2011). It is a method that is loosely motivated by the appraisal theory of emotion (i.e., emotion is a result of an individual’s cognitive assessment), which is theorized to be in stages, of a stimulus. This theory inspires a computational framework of emotion recognition in which the method is based on first processing the clear perceptual differences of emotion information in the acoustic features at the top (root) of the tree, and highly ambiguous emotions are recognized at the leaves of the tree.



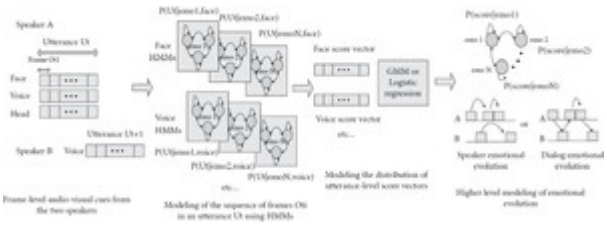
Click to view larger

Fig. 12.4 Hierarchical tree structure for multiclass emotion recognition proposed by Lee et al. (2011). The tree is composed of a binary classifier at each node; the design of the tree takes into account of emotionally relevant discrimination given acoustic behavioral cues to optimize prediction accuracy.

The key idea is that the levels in the tree are designed to solve the easiest classification tasks first, allowing us to mitigate error propagation (Figure 12.4). Each node of a tree can be a binary classifier in which the top level is designed to classify between sets of emotion classes that are most easily discriminated through modeling acoustic behaviors (e.g., angry versus sad). The leaves of the tree can be used to identify the most ambiguous emotion class, which often is the class of *neutral*. The framework was evaluated on two different emotional databases using audio-only features, the FAU AIBO database and the USC IEMOCAP database. In the FAU AIBO database, it obtained a balanced recall on each of the individual emotion classes, and the performance measure improves by 3.37% absolute (8.82% relative) over using a standard support vector machine baseline model. In the USC IEMOCAP database, it achieved an absolute (p. 179) improvement of 7.44% (14.58%) also over a baseline support vector machine modeling.

Context-Sensitive Emotion Recognition in Spoken Dialogues

In human-human interaction, the emotion of each interaction participant is temporally smooth and conditioned on the emotion state on the other speaker. Such conditional dependency between the two interacting partners' emotion states and their own temporal dynamics in a dialogue has been explicitly modeled, for example, using a dynamic Bayesian network (Lee, Busso, Lee, & Narayanan, 2009). Lee et al. applied the framework to recognizing emotion attributes described using a valence-activation dimension with speech acoustic features. Results showed improvements in classification accuracy by 3.67% absolute and 7.12% relative over the Gaussian mixture model (GMM) baseline on isolated turn-by-turn (static) emotion classification for the USC IEMOCAP database.



Click to view larger

Fig. 12.5 Context sensitive emotion recognition. Metallinou et al. proposed a flexible context-sensitive emotion recognition framework that captures both the utterance-level emotional dynamics and the long-range context dependencies of emotional flow in dialogues.

Sources: Metallinou, Katsamanis, et al. (2012) and Metallinou, Wöllmer, et al. (2012).

Other studies have examined different modeling techniques in a more general setup of context-sensitive framework (i.e., modeling emotions between interlocutors' emotion in a given dialogue (Mariooryad & Busso, 2013; Metallinou, Katsamanis, et al., 2012; Metallinou, Wöllmer, et al., 2012; Wöllmer et al., 2008; Wöllmer, Kaiser, Eyben, Schuller, & Rigoll, 2012). In particular, Metallinou et al. (Metallinou, Katsamanis, et al., 2012; Metallinou, Wöllmer, et al., 2012) have proposed a context-sensitive emotion recognition framework (see Figure 12.5). The idea was centered on the fact that emotional content of past and future observations can offer additional contextual information benefiting the emotion classification accuracy of the current utterances. Techniques such as bidirectional long short-term memory (BLSTM) neural networks, hierarchical hidden Markov model classifiers (HMMs) and hybrid HMM/BLSTM classifiers were used for modeling emotional flow within an utterance and between utterances over the course of a dialogue. Results from these studies further underscore the importance and usefulness of jointly model interlocutors and incorporating surrounding contexts to improve recognition accuracies.

Tracking of Continuously Rated Emotion Attributes

Another line of work that has emerged recently aims at describing emotion as a continuous *flow* instead of a sequence of discrete-states (i.e., a time-continuous profile instead of one decision per speech turn). In real life, many expressive behaviors and emotion manifestations are often subtle and difficult to be assigned into discrete categories. Metallinou et al. have addressed this issue by tracking continuous levels of a participant's activation, valence, and dominance during the entire course of dyadic interactions without restriction on assigning a label just for each speaking turn (Angeliki Metallinou, Katsamanis, & Narayanan, 2012).

The computational technique is based on a Gaussian mixture model–based approach that computes a mapping from a set of observed audiovisual cues to an underlying emotional state—that is, given by annotators rating over time on a continuous scale (values range from –1 to 1) along the axis of valence, activation, and dominance. The continuous emotion annotation tool is based on Feeltrace (Cowie et al., 2000). Promising results were obtained in tracking trends of participant' activation and dominance values with the GMM-based approach (p. 180) compared to other regression-based approaches in a database of two actors' improvisations (Angeliki Metallinou, Lee, Busso, Carnicke, & Narayanan, 2010). The tracking of continuously rated emotion attributes is an area of research still in its formative stages, and attempts to complement the standard approach

of assigning a specific segment of data to predefined discrete categorical emotional attributes.

Open Challenges

Each of the aforementioned three components in the design of a reliable emotion recognizer remains an active research direction. The inherent ambiguity in emotion categorizations, the variability of acoustic features in different conditions, the complex nature of the interplay between the linguistic and paralinguistic aspects manifested in speech as well as the interplay between the speech signal and signals of visual nonbehavior, and the nature of human coupling and interaction in emotional expression and perception are some of the key issues that need deeper investigation and further advance in the related computational frameworks.

Speech in Affective Computing: Future Works and Applications

Future challenges in the area of affective computing with speech lie in both improving our understanding of emotional speech production mechanisms and in designing generalizable cross-domain robust emotion recognition systems. In summary, on the acoustic feature extraction side, while the common data processing approach of feature extraction has been able to provide the state-of-art emotion recognition accuracy, it remains unclear how exactly emotional information is encoded in these acoustic waveforms. Also, the current approaches of feature computation are often difficult to be generalized across and scaled-up to real-life applications. With growing knowledge and insights into articulatory and voice source movements and their interplay in the emotion encoding process, the related acoustic feature extraction procedure in the acoustic domain can be further advanced. This holds promises to a more robust and principled ways for speech emotion processing.

Another hurdle in affective computing is the ability to obtain reliable cross-domain (and cross-corpora) recognition results. Until now, most of the emotion recognition efforts have concentrated on optimizing recognition accuracy for an individual database. Few works have started to examine the technique to achieve higher accuracy across corpora (Bone, Lee, & Narayanan, 2012; Schuller et al., 2010). It is inherently a much more difficult modeling task on top of the issues that one has to solve related to the subjectivity in the design of the emotional attributes, the lack of solid understanding on which acoustic features are robust across databases, and the issue of modeling the interactive nature of human affective dynamics. All of these remain as open questions to be

investigated in paving the way for robust real-life emotion recognition engineering systems of the future.

Having the ability to infer a person's emotional state from speech is of great importance to many scientific domains. This is because emotion is a fundamental attribute governing the generation of human expressive behavior and a key indicator in developing human behavior analytics and in designing novel user interfaces for a wide range of disciplines. Exemplary domains for such applications include commerce (e.g., measuring user frustration and satisfaction), medicine (e.g., diagnosis and treatment), psychotherapy (e.g., tracking in distressed couples research, addiction, autism spectrum disorder, depression, posttraumatic stress disorder), and educational settings (e.g., measuring engagement). Affective computing is indeed an integral component and a key building block in the field of behavioral signal processing.

Acknowledgments

The authors would like to thank National Institute of Health, National Science Foundation, United States Army, and Defense Advanced Research Agency for their funding support.

References

Bachorowski, J.-A. (1999). Vocal expression and perception of emotion. *Current Directions in Psychological Science*, *8*, 53–57.

Bitouk, D., Verma, R., & Nenkova, A. (2010). Class-level spectral features for emotion recognition. *Speech Communication*, *52*(7–8), 613–625. doi:10.1016/j.specom.2010.02.010

Black, M. P., Georgiou, P. G., Katsamanis, A., Baucom, B. R., & Narayanan, S. (2011). "You made me do it": Classification of blame in married couples' interactions by fusing automatically derived speech and language information. *Proceedings of interspeech* (pp. 89–92).

Black, M. P., Katsamanis, A., Baucom, B. R., Lee, C.-C., Lammert, A. C., Christensen, A., Georgiou, P. G.,…& (p. 181) Narayanan, S. (2013). Toward automating a human behavioral coding system for married couples' interactions using speech acoustic features. *Speech Communication*, *55*, 1–21.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, *5*, 341–345.

Bone, D., Lee, C.-C., & Narayanan, S. (2012). A robust unsupervised arousal rating framework using prosody with cross-corpora evaluation. *Proceedings of Interspeech*.

Bone, D., Li, M., Black, M. P., & Narayanan, S. S. (2014). Intoxicated speech detection: A fusion framework with speaker-normalized hierarchical functionals and GMM supervectors. *Computer Speech & Language*, *28*:2, 375–391

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

Busso, C., Lee, S., & Narayanan, S. S. (2009). Analysis of emotionally salient aspects of fundamental frequency for emotion detection. *IEEE transactions on audio, speech and language processing*, *17*, 582–596. doi:10.1109/TASL.2008.2009578

Busso, C., Metallinou, A., & Narayanan, S. (2011). Iterative feature normalization for emotional speech detection. *international conference on acoustics, speech, and signal processing (ICASSP)* (pp. 5692–5695).

Busso, C., & Narayanan, S. S. (2008). Recording audio-visual emotional databases from actors: a closer look. *Second international workshop on emotion: Corpora for research on emotion and affect, international conference on language resources and evaluation (LREC 2008)* (pp. 17–22), Marrakesh, Morocco.

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *2, 27*, 1–27.

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

Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., & Schröder, M. (2000). FEELTRACE: An instrument for recording perceived emotion in real time. *ISCA tutorial and research workshop (ITRW) on speech and emotion* (pp. 19–24). Winona, MN: International Society for Computers and Their Applications.

Deller, J. R., Hansen, J. H. L., & Proakis, J. G. (2000). *Discrete-time processing of speech signals*. Piscataway, NJ: IEEE Press.

Erickson, D., Menezes, C., & Fujino, A. (2004). Some articulatory measurements of real sadness. *Proceedings of interspeech* (pp. 1825–1828).

Eyben, F., Wöllmer, M., & Schuller, B. (2010). OpenSMILE: The Munich versatile and fast open-source audio feature extractor. *ACM international conference on multimedia (MM 2010)* (pp. 1459–1462).

Fant, G. (1970). *Acoustic theory of speech production*. The Hague, Netherlands: Walter de Gruyter.

Fujimura, O., Kiritani, S., & Ishida, H. (1973). Computer controlled radiography for observation of movements of articulatory and other human organs. *Computers in Biology and Medicine*, *3*, 371–384.

Gobl, C., & Chasaide, A. N. (2003). The role of voice quality in communicating emotion, mood and attitude. *Speech Communication—Special issue on speech and emotion*, *40*, 189–212.

Hall, M. A. (1999). *Correlation-based feature selection for machine learning*. Hamilton, New Zealand: The University of Waikato.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter*, *11*, 10–18. doi:10.1145/1656274.1656278

Jain, A., & Zongker, D. (1997). Feature selection: Evaluation, application, and small sample performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*, 153–158.

Juslin, P. N., & Scherer, K. R. (2005). Vocal expression of affect. *The new handbook of methods in nonverbal behavior research*, 65–135, New York City, New York: Oxford University Press

Kazamzadeh, A., Lee, S., Georgiou, P., & Narayanan, S. (2011). Emotion twenty question (EMO20Q): Toward a crowd-sourced theory of emotions. *Proceedings of affective computing and intelligent interaction (ACII)* (pp. 1–10), Memphis, Tennessee

Kim, J., Lee, S., & Narayanan, S. (2010). A study of interplay between articulatory movement and prosodic characteristics in emotional speech production. *Proceedings of interspeech* (pp. 1173–1176).

Kim, J., Lee, S., & Narayanan, S. (2011). An exploratory study of the relations between perceived emotion strength and articulatory kinematics. *Proceedings of interspeech* (pp. 2961–2964).

Kipp, M. (2001). ANVIL—A generic annotation tool for multimodal dialogue. *European conference on speech communication and technology (Eurospeech)* (pp. 1367–1370).

Lammert, A., Proctor, M., & Narayanan, S. (2013). Morphological variation in the adult hard palate and posterior pharyngeal wall. *Speech, Language, and Hearing Research*, *56*, 521–530.

Le, X., Quénot, G., & Castelli, E. (2004). Recognizing emotions for the audio-visual document indexing. *Ninth international symposium on computers and communications (ISCC)* (Vol. *2*, pp. 580–584).

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

Lee, C.-C., Busso, C., Lee, S., & Narayanan, S. S. (2009). Modeling mutual influence of interlocutor emotion states in dyadic spoken interactions. *Proceedings of interspeech* (pp. 1983–1986).

Lee, C.-C., Mower, E., Busso, C., Lee, S., & Narayanan, S. S. (2011). Emotion recognition using a hierarchical binary decision tree approach. *Speech Communication*, *53*, 1162–1171. doi:10.1016/j.specom.2011.06.004

Lee, S., Bresch, E., Adams, J., Kazemzadeh, A., & Narayanan, S. S. (2006). A study of emotional speech articulation using a fast magnetic resonance imaging technique. *International Conference on spoken language (ICSLP)* (pp. 2234–2237).

Lee, S., Yildirim, S., Kazemzadeh, A., & Narayanan, S. S. (2005). An articulatory study of emotional speech production. *Proceedings of Interspeech* (pp. 497–500).

Mariooryad, S., & Busso, C. (2013). Exploring cross-modality affective reactions for audiovisual emotion recognition. *IEEE Transactions on Affective Computing*. In press. doi: 10.1109/T-AFFC.2013.11

Metallinou, A., Katsamanis, A., & Narayanan, S. S. (2012). A hierarchical framework for modeling multimodality and emotional evolution in affective dialogs. *International* *(p. 182)* *conference on acoustics, speech, and signal processing (ICASSP)* (pp. 2401–2404). doi:10.1109/ICASSP.2012.6288399

Metallinou, A., Wöllmer, M., Katsamanis, A., Eyben, F., Schuller, B., & Narayanan, S. S. (2012). Context-sensitive learning for enhanced audiovisual emotion classification. *IEEE Transactions on Affective Computing*, *3*, 184–198. doi:10.1109/T-AFFC.2011.40

Metallinou, A., Katsamanis, A., & Narayanan, S. (2013). Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information. *Image and Vision Computing*, *31*:2, 137–152

Metallinou, A., Lee, C.-C., Busso, C., Carnicke, S., & Narayanan, S. S. (2010). The USC CreativeIT database: a multimodal database of theatrical improvisation. *Proceedings of the multimodal corpora workshop: advances in capturing, coding and analyzing, multimodality (MMC)* (pp. 64–68), Valetta, Malta

Mower, E., Mataric, M. J., & Narayanan, S. S. (2011). A framework for automatic human emotion classification using emotional profiles. *IEEE Transactions on Audio, Speech and Language Processing*, *19*:5, 1057–1070.

Mower, E., Metallinou, A., Lee, C.-C., Kazemzadeh, A., Busso, C., Lee, S., & Narayanan, S. (2009). Interpreting ambiguous emotional expressions. *Proceedings of affective computing and intelligent interaction and workshops (ACII)* (pp. 1–8), Amsterdam, Netherlands

Murphy, P. J., & Laukkanen, A.-M. (2009). Electroglottogram analysis of emotionally styled phonation. *Multimodal signals: Cognitive and algorithmic issues*, 264–270, Vietri sul Mare, Italy

Narayanan, S. S., Alwan, A. A., & Haker, K. (1995). An articulatory study of fricative consonants using magnetic resonance imaging. *The Journal of the Acoustical Society of America*, *98*, 1325–1347.

Narayanan, S. S., & Georgiou, P. (2013). Behavioral signal processing: Deriving human behavioral informatics from speech and language. *Proceedings of the IEEE*, *101*, 1203–1233. doi:10.1109/JPROC.2012.2236291

Narayanan, S. S., Nayak, K., Lee, S., Sethy, A., & Byrd, D. (2004). An approach to real-time magnetic resonance imaging for speech production. *The Journal of the Acoustical Society of America*, *115*, 1771–1776.

Pao, T.-L., Yeh, J.-H., Chen, Y.-T., Cheng, Y.-M., & Lin, Y.-Y. (2007). A comparative study of different weighting schemes on knn-based emotion recognition in Mandarin speech. In D.-S. Huang, L. Heutte, & M. Loog (Eds.), *advanced intelligent computing theories and applications with aspects of theoretical and methodological issues* (pp. 997–1005). Berlin: Springer-Verlag. doi:10.1007/978-3-540-74171-8_101

Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*, 1226–1238.

Perkell, J. S., Cohen, M. H., Svirsky, M. A., Matthies, M. L., Garabieta, I., & Jackson, M. T. (1992). Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements. *The Journal of the Acoustical Society of America*, *92*, 3078–3096.

Rahman, T., & Busso, C. (2012). A personalized emotion recognition system using an unsupervised feature adaptation scheme. *International conference on acoustics, speech, and signal processing (ICASSP)* (pp. 5117–5120).

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech communication*, *40*, 227–256.

Schuller, B., Arsic, D., Wallhoff, F., & Rigoll, G. (2006). Emotion recognition in the noise applying large acoustic feature sets. *Speech prosody*, (pp. 276–289), Dresden, Germany.

Schuller, B., Batliner, A., Seppi, D., Steidl, S., Vogt, T., Wagner, J., Devillers, L.,…& VeredAharonson (2007). The relevance of feature type for the automatic classification of emotional user states: Low level descriptors and functionals. *Proceedings of interspeech* (pp. 2253–2256).

Schuller, B., Rigoll, G., & Lang, M. (2003). Hidden Markov model–based speech emotion recognition. *International conference on acoustics, speech, and signal processing (ICASSP)* (Vol. *2*, pp. 1–4).

Schuller, B., Vlasenko, B., Eyben, F., Wöllmer, M., Stuhlsatz, A., Wendemuth, A., & Rigoll, G. (2010). Cross-corpus acoustic emotion recognition: Variances and strategies. *IEEE Transactions on affective computing*, *1*(2), 119–131. doi:10.1109/T-AFFC.2010.8

Schuller, B., Vlasenko, B., Minguez, R., Rigoll, G., & Wendemuth, A. (2007). Comparing one and two-stage acoustic modeling in the recognition of emotion in speech. *IEEE workshop on automatic speech recognition & understanding (ASRU)* (pp. 596–600).

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., & Narayanan, S. S. (2013). Paralinguistics in speech and language—State-of-the-art and the challenge. *Computer Speech & Language*, *27*, 4–39.

Sethu, V., Ambikairajah, E., & Epps, J. (2007). Speaker normalisation for speech based emotion detection. *15th International Conference on Digital Signal Processing (DSP)* (pp. 611–614).

Stone, M. (2005). A guide to analysing tongue motion from ultrasound images. *Clinical Linguistics & Phonetics*, *19*, 455–501.

Vlasenko, B., Schuller, B., Wendemuth, A., & Rigoll, G. (2007). Frame vs. turn-level: Emotion recognition from speech considering static and dynamic processing. In A. Paiva, R. Prada, & R. W. Picard (Eds.), *Affective computing and intelligent interaction* (pp. 139–147). Berlin and Heidelberg: Springer. doi:10.1007/978-3-540-74889-2_13

Watson, D., Clark, L. A., & Tellegen, A. (1988). Developement and validation of brief measures of positive and negative affect: The PANAS Scale. *Personality and Social Psychology*, *47*, 1063–1070.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). Elan: A professional framework for multimodality research. *Proceedings of LREC* (Vol. 2006), Genoa, Italy

Wöllmer, M., Eyben, F., Reiter, S., Schuller, B., Cox, C., Douglas-Cowie, E., & Cowie, R. (2008). Abandoning emotion classes—Towards continuous emotion recognition with modelling of long-range dependencies. *Proceedings of Interspeech* (pp. 597–600)..

Wöllmer, M., Kaiser, M., Eyben, F., Schuller, B., & Rigoll, G. (2012). LSTM-Modeling of continuous emotions in an audiovisual affect recognition framework. *Image and Vision Computing*, *31*(2), 153–163. doi:10.1016/j.imavis.2012.03.001

(p. 183) Yan, Z., Li, Z., Cairong, Z., & Yinhua, Y. (2008). Speech emotion recognition using modified quadratic discrimination function. *Journal of Electronics (China)*, *25*, 840–844. doi:10.1007/s11767-008-0041-8

Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., Busso, C., Deng, Z., Lee, S., et al. (2004). An acoustic study of emotions expressed in speech. *International conference on spoken language processing (ICSLP)* (pp. 2193–2196).

Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D.,…& Woodland, P. (2002). The HTK book. Cambridge University Engineering Department, *3*, 175.

Zeng, Z., Pantic, M., Roisman, G. I., & Huang, T. S. (2009). A survey of affect recognition methods: audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence*, *31*, 39–58.

Notes:

(1) . http://sourceforge.net/projects/wavesurfer

(2) . http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

Chi-Chun Lee

Viterbi School of Engineering, University of Southern California, Los Angeles, CA

Jangwon Kim

Viterbi School of Engineering, University of Southern California, Los Angeles, CA

Angeliki Metallinou

Viterbi School of Engineering, University of Southern California, Los Angeles, CA

Carlos Busso

Carlos Busso, Electrical Engineering Department, The University of Texas at Dallas, TX

Sungbok Lee

Sungbok Lee, Viterbi School of Engineering, University of Southern California, Los Angeles, CA

Shrikanth S. Narayanan

Shrikanth S. Narayanan, Viterbi School of Engineering, University of Southern California, Los Angeles, CA