Emotion Recognition from Speech: A Survey

Georgios Drakopoulos, George Pikramenos, Evaggelos Spyrou and Stavros J. Perantonis
NCSR “Demokritos”, Athens, Greece

Keywords: Emotion Recognition, Affective Computing, Signal Processing, Cepstrum, MFCC Coefficients, Deep Neural Networks, Time Series Analysis, Hidden Markov Chains, Higher Order Data.

Abstract: Emotion recognition from speech signals is an important field in its own right as well as a mainstay of many multimodal sentiment analysis systems. The latter may as well include a broad spectrum of modalities which are strongly associated with consciously or subconsciously communicating human emotional state such as visual cues, gestures, body postures, gait, or facial expressions. Typically, emotion discovery from speech signals not only requires considerably less computational complexity than other modalities, but also at the same time in the overwhelming majority of studies the inclusion of speech modality increases the accuracy of the overall emotion estimation process. The principal algorithmic cornerstones of emotion estimation from speech signals are Hidden Markov Models, time series modeling, cepstrum processing, and deep learning methodologies, the latter two being prime examples of higher order data processing. Additionally, the most known datasets which serve as emotion recognition benchmarks are described.

1 INTRODUCTION

Emotion discovery from speech offers a unique tool for estimating with considerable accuracy human affective state with remarkably low computationally complexity, especially when compared with the task of human activity discovery based on video. Affective states reveal the subjective understanding of an individual of an external stimulus or condition and, moreover, may well serve as intention indications, since emotions are the true driving force behind many human actions or reactions. Additionally, affective state estimation plays an important role in cognitive sciences (Cowie et al., 2001) and affective computing (Picard, 2003)(Tao and Tan, 2005).

The primary objective of this conference paper is the identification of the main research pylons in the field of emotion discovery from speech, to illustrate the differences between them, and to present some of the main scientific literature works underlying each such pylon. Moreover, as a secondary objective, some of the most popular online multimodal datasets which include speech are presented.

The remainder of this work is structured as follows. The recent scientific literature is briefly reviewed in section 2. The primary methodological frameworks for affective state estimation are presented in section 3. In section 4 the major public datasets concerning emotion recognition from speech are described, whereas in section 5 future research directions are explored. Finally, in table 1 the nnotation of this conference paper is summarized.

Table 1: Paper Notation.

Symbol	Meaning
$\triangleq$	Equality by definition
$\{s_1,\ldots,s_n\}$	Set with elements $s_1,\ldots,s_n$
$\mathrm{card}\left(S\right)$	Set cardinality
$\left\|X\left(e^{j\omega}\right)\right\|$	Magnitude of $X\left(e^{j\omega}\right)$
$\mathcal{F}\left[x\left(t\right)\right]$	Fourier transform of $x\left(t\right)$
$\mathcal{F}^{-1}\left[X\left(e^{j\omega}\right)\right]$	Inverse FT of $X\left(e^{j\omega}\right)$
$\langle x_1\left(t\right) \mid x_2\left(t\right)\rangle$	Inner product of $x_1\left(t\right)$ and $x_2\left(t\right)$

2 PREVIOUS WORK

Emotion recognition has taken many forms in scientific literature, both as part of the broader humanistic data mining field but also on its own right (Kwon et al., 2003). Many engineering approaches for emotion discovery follow a black box approach and rely on observing emotional traits in such as in facial expressions as in (Goldman and Sripada, 2005)(Haxby et al., 2002). Moreover, multifactor facial analysis with tensor classification (Vasilescu and Terzopoulos,

Drakopoulos, G., Pikramenos, G., Spyrou, E. and Perantonis, S.
Emotion Recognition from Speech: A Survey.
DOI: 10.5220/0008495004320439
In *Proceedings of the 15th International Conference on Web Information Systems and Technologies (WEBIST 2019)*, pages 432-439
ISBN: 978-989-758-386-5

2002)(Tian et al., 2012) or tensor subspace (Cai et al., 2005)(Cichocki et al., 2015) algorithms have been explored. Inclusion of more traits has led to bimodal (De Silva and Ng, 2000) and multimodal (Busso et al., 2004)(Kim et al., 2004) emotion estimators based on various physiological signs.

Alternatively, a plethora of white box approaches have been also proposed. Patterns in ECG (Agrafioti et al., 2012) or EEG waveforms (Mohammadi et al., 2017)(Murugappan et al., 2010) have been used to deduce affective states. This can be also achieved indirectly with the BCI proposed in (Mathe and Spyrou, 2016) as part of an IoT ecosystem.

Concerning human activity and creation, art and especially music have close ties with the underlying affective states (Li and Ogihara, 2004). Emotion discovery in music is the focus of (Busso et al., 2009). The affective reults of music are explored in (Kim and André, 2008), (Lin et al., 2010), and (Yang et al., 2008) which proposes affective categorical regression with arousal and valence as input variables. The effects of acoustic features such as jitter and shimmer are evaluated in (Li et al., 2007) and (Jin et al., 2015). Finally, (Elfenbein and Ambady, 2002) is a meta-analysis of the connection between affective states and music based on cultural factors.

Affective states can also play a central role in online social network analysis and specifically in information diffusion and digital influence. Applying a modified version of the methodology proposed in (Drakopoulos et al., 2017b) for an extension of term-document matrix to a term-keyword-document third order tensor, (Drakopoulos et al., 2017a) formulates higher order influence metrics which can be easily extended to include affective information. The same methodology can be applied to assess the compactness of spatio-linguistic online communities as discovered in (Drakopoulos et al., 2019) or to communities defined by multiple interaction paths (Drakopoulos, 2016).

The primary emotions according to the groundbreaking theory developed by Plutchik, as stated among others in (Plutchik et al., 1979), (Plutchik, 1980), and (Plutchik, 2001), are joy, trust, expectation, fear, sadness, disgust, anger, and surprise (Wallbott and Scherer, 1986). Based on this theory, secondary and tertiary emotions can be derived by superimposing the above primary emotions in various scales (Lane et al., 1996). For instance, under this model love is derived as the composition of joy and trust. Moreover, it should also be noted that, even though it is not an emotion per se, the netural state is a valid emotional state. Note that in other contexts other emotions may well form the basis for research (Kohler et al., 2000). One such prominent case is education, where pride, remorse, boredom, or guilt are sought to be invoked or detectted during teaching activity (Jerritta et al., 2011)(Spyrou et al., 2018).

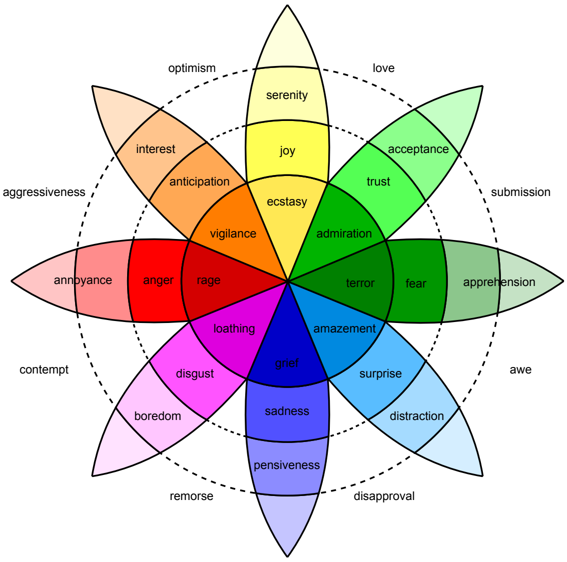

Figure 1: Emotion wheel (from (Plutchik, 2001)).

3 EMOTION RECOGNITION FROM SPEECH

3.1 Hidden Markov Models

Hidden Markov Models (HMMs) are almost synonymous with speech processing and they come in two flavors, depending on the amount of observable information known to the researcher (El Ayadi et al., 2011). Let us S denote the state set and $n = \operatorname{card}(S)$ be its cardinality:

$$S \triangleq \{s_1, \ldots, s_n\} \tag{1}$$

Additionally, for each state $s_i \in S$ let S_i be the set of each possible outbound transitions from s_i in one step:

$$S_i \triangleq \left\{ s_i \rightarrow s_j, s_j \in S \right\}, \qquad 1 \leq i \leq n \tag{2}$$

Finally, let P contain the individual transition probabilities as:

$$P \triangleq \left\{ \text{prob}\left\{ s_i \rightarrow s_j \right\}, \forall s_i, s_j \in S \right\} \tag{3}$$

The two variants of HMMs are:

- S and S_i are known and the elements of P must be estimated, usually by statistical methods including classical or Bayesian estimation.

- S, and by extension S_i, are unknown. The cardinality $\operatorname{card}(S)$ may be have to be estimated as well, depending on the problem. In this case the sets S, S_i, and P must be estimated only from the output symbols. Typically this is achieved with mutual information or other divergence metrics (Bahl et al., 1986).

In (Schuller et al., 2003) two methodologies for estimating the parameters of an HMM corresponding to six basic emotional states are presented. The first is a mixture of Gaussians model whose local maxima are functions of said states weighted by the the probabilities of a number of features including pitch-related statistics such as its location and its maximum absolute deviation in the voice sample. The second way is based on embedding local emotional distributions and computing their maximum using the same set of features. Gaussian mixture models are also considered in (Li et al., 2013a), but in conjunction with restricted Boltzmann models. For a versatile and persistent data structure which can represent HMMs with a variable number of states see (Kontopoulos and Drakopoulos, 2014).

HMMs act as classifiers among the archetypal emotions of anger, disgust, fear, joy, sadness, and surprise in (Nwe et al., 2003), where speech signals are decomposed in its short time log frequency power coefficients. The selection of these particular features leads to improved accuracy compared to representations based on fundamental frequency, the ratio between silence and speech, and energy contour.

3.2 Deep Learning

Neural networks of different configurations acting as affective classifiers have been proposed. For instance, (Kobayashi and Hara, 1992) considers feedforward neural networks, (Sprengelmeyer et al., 1998) and (Adolphs, 2002) explore the connection of emotion recognition through different human neural substrates based on findings from neurophysiological disorders, and (Nicholson et al., 2000) considers a neural network trained by phoneme balanced words to distinguish between eight emotional states. Estimating emotion from speech is also the goal of the neural network architecture proposed in (Bhatti et al., 2004). This is extended to deep neural networks with multiple hidden layers and a combination of activation functions in (Stuhlsatz et al., 2011). Multimodal architectures based on facial expressions and speech are the focus of (Ioannou et al., 2005) and (Kahou et al., 2013). Moreover, (Lin and Wei, 2005) suggests using a Support Vector Machine (SVM) over HMM in order to increase the accuracy of recognizing basic emotional states from speech signals. SVMs are also considered in the context of a smart home ecosystem where distributed IoT processing can assess the affective state of its inhabitants and adapt accordingly (Pan et al., 2012).

Extreme learning machines (ELMs) proposed among others in (Han et al., 2014) have a simple architecture comprising of one single long hidden layer of neurons with non-linear activation functions. This allows for closed form expressions connecting ELM input and output to be constructed (Huang et al., 2006), which in turn leads to an easy and controllable training process. Specifically:

- The k-th input neuron forwards the k-th component x_k of the current training vector.
- The i-th hidden neuron receives $w^0_{k,i}x_k$, sums each such stimulus and then subtracts threshold β_i, and drives the result to its non-linear activation function $\psi(\cdot)$.
- The j-th output neuron y_j repeats this process with its own synaptic weights $w^1_{i,j}$, non-linear activation function $\varphi(\cdot)$, and threshold β_j to generate:

$$y_j \triangleq \varphi\left(\sum_{i\to j} w^1_{i,j}\psi\left(\sum_{k\to i} w^0_{k,i}x_k - \beta_i\right) - \beta_j\right) \tag{4}$$

3.3 Signal Processing

Signal processing of a speech signal either in its original time domain waveform $x(t)$ or in a number of various transformations can yield important information regarding the speaker emotional state. Regarding time domain, a statistical method for distinguishing between joy, anger, sadness, fear, and neutral emotional state based on speech signal characteristics such as maximum pitch and maximum absolute pitch divergence is presented in (Petrushin, 2000).

In (Wu et al., 2011) an elaborate filter bank based on modulations and Hilbert envelope is proposed in order to extract from speech auditory-inspired long-term spectro-temporal features which contains vital temporal and spectral acoustic information, without resorting to short-term features. The study of cepstrum $\tilde{x}(t)$, and in particular of its MFCC coefficients, of the speech signal $x(t)$ appears in many works. Recall that:

$$\tilde{x}(t) \triangleq \left|\mathcal{F}^{-1}\left[\ln|\mathcal{F}[x(t)]|^2\right]\right|^2 \tag{5}$$

The amplitude of the coefficients of the discrete cosine transform of $\tilde{x}(t)$ are the cepstral or MFCC coefficients of $x(t)$.

Power cepstrum coefficients are used in (Sato and Obuchi, 2007) instead of prosodic features in order to express the speech signal in scales which are closer to that of human audio perception. Specifically, phonetic features, expressed in the form of cepstral coefficients increase classification accuracy over over an utterance. Along a similar line of reasoning, in (Dumouchel et al., 2009) the cepstral coefficients of Gaussian mixture models are used in order to discern between basic emotions.

Another approach is the wavelet transform which reveals information about $x(t)$ in multiple time scales, potentially discovering more patterns than thse of the classical Fourier transform (Daubechies, 1990)(Antonini et al., 1992). The central idea of the wavelet transform is to use a family of basis functions indexed by location μ_0 and scale σ_0 parameters. Although there is a plethora of wavelet basis such as the Morlet, the log-Gabor, or the Haar families, perhaps the most common example is the Gaussian kernel family:

$$g(t;\mu_0,\sigma_0) \overset{\triangle}{=} \frac{1}{\sigma_0\sqrt{2\pi}} \exp\left(-\frac{(t-\mu_0)^2}{2\sigma_0^2}\right) \tag{6}$$

The projection of $x(t)$ to various members of the wavelet basis family yields the wavelet coefficients:

$$\begin{aligned} w_{\mu_0,\sigma_0} &\overset{\triangle}{=} \langle x(t) \mid g(t;\mu_0,\sigma_0)\rangle \\ &= \int_\Omega x(t)\, g(t;\mu_0,\sigma_0)\, dt \end{aligned} \tag{7}$$

Wavelet transform is combined with cepstral coefficients and the subbabd based cepstral parameter in (Kishore and Satish, 2013) for affective state estimation. Furthermore, multimodal emotion recognition from video and speech through wavelets is described in (Go et al., 2003), with the note that the addition of speech to video increases classification accuracy.

4 DATASETS

4.1 Audio Datasets

- Perhaps the most well known English dataset is the Toronto Emotional Speech Set (Dupuis and Pichora-Fuller, 2010) which contains audio only data collected at Northwestern University according to the Auditory Test Protocol no. 6 of the same university. Two professional female actresses, one of 26 and one of 64 years old, born and raised in Toronto area uttered 2800 words corresponding to seven basic emotional states.
- RAVDESS (Livingstone et al., 2012) is also a popular dataset consisting of 7356 files, each of which has been evaluated 247 times by an equal number of North American volunteers in terms of sentimental validity and intensity. Moreover, 72 additional participants evaluated the same data based on a test-retest methodology. Both the behavior of each individual evaluator and the evaluations themselves were remarkably consistent.
- The Emo-Soundscapes collection (Fan et al., 2017) contains two benchmark protocols as well as 613 sound clips coming from a combination of 600 music files from www.freesound.org. Along with the original files there are 1213 music excerpts under the Creative Commons license. Each such excerpt lasts six seconds and its induced emotional intensity was evaluated through crowdsourcing by 1182 people from 74 countries around the globe.
- The Speech Under Simulated and Actual Stress (SUSAS) (Hansen and Bou-Ghazale, 1997) was created by University of Colorado-Boulder and the US Air Force Research Laboratory. SUSAS contains more than 16000 emotionally charged sentences spoken under various stress levels from 32 speakers (13 men and 19 women) of ages between 22 and 76. Moreover, SUSAS includes large files with communications from four Apache helicopter pilots with the control tower along with their transcripts from the Linguistic Data Consortium under the name SUSAS Transcripts (LDC99T33).

Emotionally charged speech datasets are available for a number of other languages as well, most of which belong to the Indo-European language family. Such datasets allow researchers to exclude linguistic or cultural factors from the discovery process, focusing only on the emotional content.

- The FAU Aibo emotion corpus (Batliner et al., 2008) has been created from the use of Aibo, a Sony pet robot, from 51 German children aged between 10 and 13 years old. This corpus contains spontaneous and emotionally charged commands which have been collected and broken down to elementary parts based on syntax and prosody, which in turn are manually assigned one out of 11 possible emotional labels from five evaluators.
- A second dataset is BAUM-1 (Zhalehpour et al., 2016) made up of short video clips, each approximately four minutes long. In each such clip German male professional actors utter sentences corresponding to a plethora of emotional states including joy, anger, sadness, disgust, fear, surprise, confusion, disdain, and annoyance as well as the neutral state.

- A smaller German dataset is the Berlin database of emotional speech (Burkhardt et al., 2005) which contains 500 samples of speech uttered by ten professional actors. Each such sample can represent six emotional states.
- The RML emotion dataset (Wang and Guan, 2008) from Ryerson Multimedia Lab consists of 720 audio-visual clips, each lasting from 3 to 6 seconds and with a single emotional charge out of six possible basic emotional states. These are at least ten video clips for anger, disgust, fear, joy, sadness, and surprise. In order for the eight volunteers to utter sentences which genuinely contain a given emotion, they were asked to recall an indicent from their own lives which caused that emotion. Moreover, they are native speakers of either English, Mandarin Chinese, Farsi, Italian, Urdu, or Panjabi.

4.2 Multimodal Datasets

In contrast to audio-only datasets, multimodal ones allow the separate examination of how isolated modalities like speech, text, facial expressions, gait, or body posture or a combination thereof are involved in emotion discovery.

- CREMA-D (Cao et al., 2014) relies on the human trait of communicating emotional state through voice and facial expression. To this end, the dataset consists of video clips with speech and facial expressions covering the spectrum of basic emotions, namely joy, sadness, anger, fear, disgust, and neutral state from 91 actors of various nationalities. Each such clip is independently labeled regarding the emotion and its intensity through crowdsourcing by 2443 evaluators based on either one of the modalities or both. According to these evaluations, the neutral state was the easiest to identify, followed by joy, anger, disgust, fear, and sadness.
- The British Surrey Audio-Visual Expressed Emotion (SAVEE) dataset (Jackson and Haq, 2014) contains 480 video clips of four British male actors in seven emotional states. The sentences corresponding to these states were selected from the TIMIT corpus so that each state is equally represented. Ten human evaluators estimated the state to create the ground truth for classification algorithms.
- The Interactive Emotional Dyadic Motion Capture (IEMOCAP) (Busso et al., 2008) is a multimodal dataset maintained by USC SAIL Lab based on approximately twelve hours in total of audio-visual clips enriched with text from multiple authors. In each such clip is recorded a meeting between two actors, who can either act based on a script or can improvise in order to induce specific emotional reactions. The IEMOCAP clips have also been annotated by multiple reviewers in terms of a quadruple containing an emotional state, valence, activation, and dominance.
- The Oulou-CASIA NIR and VIS facial expression database (Li et al., 2013b) is made up of high resolution images of expressions corresponding to six emotional states from 80 actors. Said images were obtained from two imaging systems, one operating in the visible light spectrum (VIS) and one in the near infrared spectrum (NIR). Three typical lighting settings were used, namely normal office lighting, weak lighting coming only from computer monitors, and no lighting.
- eNTERFACE ‘05 (Martin et al., 2006) is an audio-visual dataset, where emotional content can be found in the audio modality, the visual modality, or both, depending on the case. This allows the benchmarking of machine learning algorithms which rely on either one or both modalities. Additionally, eNTERFACE includes the assumptions underlying its functionality, the challenges during its implementations, and how these were eventually addressed.
- Finally, the MSP-Improv corpus (Busso et al., 2017) contains audio-visual recordings of two-person improvisations out of a pool of twelve professional actors. These were specifically designed to cause genuine emotional reactions, on the condition that speech and facial expressions convey different emotions. In this way the factors involved in emotion recognition, a cognitive function known as recombination which relies on all senses and is central to the creation of the final stimulus.

5 CONCLUSIONS

The focus of this conference paper is twofold. First, the basic methodological schemes from emotion discovery are enumerated. Second, the most popular audio-only and multimodal datasets which contain emotional states or estimations thereof are presented. The latter serve as benchmarks for evaluating the algorithmic performance of emotion estimation techniques, primarily in terms of accuracy and scalability.

Regarding emotion discovery, the advent of deep learning algorithms is promising in the sense that not only such algorithms can efficiently handle 6V datasets, but also they can extract non-trivial knowledge from them, with the latter being considerably more concise and structured compated to the original datasets. Moreover, cross-domain knowledge transfer methodologies can be used to augment the knowledge body of a given domain with external elements. Finally, ontologies or knowledge graphs allow the generation of formal theorems from ground truth with reasoners such as Owlready for Python or the semantic engine of Neo4j.

ACKNOWLEDGMENT

This research has been co-financed by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH – CREATE – INNOVATE (project code: T1EDK-02070).

REFERENCES

Adolphs, R. (2002). Neural systems for recognizing emotion. *Current opinion in neurobiology*, 12(2):169–177.

Agrafioti, F., Hatzinakos, D., and Anderson, A. K. (2012). ECG pattern analysis for emotion detection. *IEEE Transactions on Affective Computing*, 3(1):102–115.

Antonini, M., Barlaud, M., Mathieu, P., and Daubechies, I. (1992). Image coding using wavelet transform. *IEEE Transactions on image processing*, 1(2):205–220.

Bahl, L. R., Brown, P. F., De Souza, P. V., and Mercer, R. L. (1986). Maximum mutual information estimation of hidden Markov model parameters for speech recognition. In *ICASSP*, volume 86, pages 49–52.

Batliner, A., Steidl, S., and Nöth, E. (2008). Releasing a thoroughly annotated and processed spontaneous emotional database: The FAU Aibo Emotion Corpus. In *Satellite Workshop of LREC*, volume 2008, page 28.

Bhatti, M. W., Wang, Y., and Guan, L. (2004). A neural network approach for human emotion recognition in speech. In *International symposium on circuits and systems*, volume 2, pages II–181. IEEE.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., and Weiss, B. (2005). A database of German emotional speech. In *Ninth European Conference on Speech Communication and Technology*.

Busso, C. et al. (2004). Analysis of emotion recognition using facial expressions, speech and multimodal information. In *International conference on multimodal interfaces*, pages 205–211. ACM.

Busso, C. et al. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335.

Busso, C., Lee, S., and Narayanan, S. (2009). Analysis of emotionally salient aspects of fundamental frequency for emotion detection. *IEEE transactions on audio, speech, and language processing*, 17(4):582–596.

Busso, C., Parthasarathy, S., Burmania, A., AbdelWahab, M., Sadoughi, N., and Mower Provost, E. (2017). MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception. *IEEE Transactions on Affective Computing*, 8(1):67–80.

Cai, D., He, X., and Han, J. (2005). Subspace learning based on tensor analysis. Technical report, UIUC.

Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., and Verma, R. (2014). CREMA-D: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390.

Cichocki, A. et al. (2015). Tensor decompositions for signal processing applications: From two-way to multiway component analysis. *IEEE Signal Processing Magazine*, 32(2):145–163.

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

Daubechies, I. (1990). The wavelet transform, time-frequency localization and signal analysis. *IEEE transactions on information theory*, 36(5):961–1005.

De Silva, L. C. and Ng, P. C. (2000). Bimodal emotion recognition. In *International conference on automatic face and gesture recognition*, pages 332–335. IEEE.

Drakopoulos, G. (2016). Tensor fusion of social structural and functional analytics over Neo4j. In *IISA*, pages 1–6. IEEE.

Drakopoulos, G. et al. (2017a). Defining and evaluating Twitter influence metrics: A higher-order approach in Neo4j. *SNAM*, 7(1):52:1–52:14.

Drakopoulos, G. et al. (2017b). Tensor-based semantically-aware topic clustering of biomedical documents. *Computation*, 5(3):34.

Drakopoulos, G. et al. (2019). A genetic algorithm for spatiosocial tensor clustering. *EVOS*, pages 1–11.

Dumouchel, P., Dehak, N., Attabi, Y., Dehak, R., and Boufaden, N. (2009). Cepstral and long-term features for emotion recognition. In *Annual conference of the International Speech Communication Association*.

Dupuis, K. and Pichora-Fuller, M. K. (2010). *Toronto Emotional Speech Set (TESS)*. University of Toronto, Psychology Department.

El Ayadi, M., Kamel, M. S., and Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587.

Elfenbein, H. A. and Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin*, 128(2):203.

Fan, J., Thorogood, M., and Pasquier, P. (2017). Emo-soundscapes: A dataset for soundscape emotion recognition. In *ACII*, pages 196–201. IEEE.

Go, H.-J., Kwak, K.-C., Lee, D.-J., and Chun, M.-G. (2003). Emotion recognition from the facial image and speech signal. In *SICE*, volume 3, pages 2890–2895. IEEE.

Goldman, A. I. and Sripada, C. S. (2005). Simulationist models of face-based emotion recognition. *Cognition*, 94(3):193–213.

Han, K., Yu, D., and Tashev, I. (2014). Speech emotion recognition using deep neural network and extreme learning machine. In *Fifteenth annual conference of the international speech communication association*.

Hansen, J. H. and Bou-Ghazale, S. E. (1997). Getting started with SUSAS: A speech under simulated and actual stress database. In *Fifth European Conference on Speech Communication and Technology*.

Haxby, J. V., Hoffman, E. A., and Gobbini, M. I. (2002). Human neural systems for face recognition and social communication. *Biological psychiatry*, 51(1):59–67.

Huang, G.-B., Zhu, Q.-Y., and Siew, C.-K. (2006). Extreme learning machine: Theory and applications. *Neurocomputing*, 70(1-3):489–501.

Ioannou, S. V. et al. (2005). Emotion recognition through facial expression analysis based on a neurofuzzy network. *Neural Networks*, 18(4):423–435.

Jackson, P. and Haq, S. (2014). Surrey audio-visual expressed emotion SAVEE database. *University of Surrey: Guildford, UK*.

Jerritta, S., Murugappan, M., Nagarajan, R., and Wan, K. (2011). Physiological signals based human emotion recognition: A review. In *International colloquium on signal processing and its applications*, pages 410–415. IEEE.

Jin, Q., Li, C., Chen, S., and Wu, H. (2015). Speech emotion recognition with acoustic and lexical features. In *ICASSP*, pages 4749–4753. IEEE.

Kahou, S. E. et al. (2013). Combining modality specific deep neural networks for emotion recognition in video. In *ICMI*, pages 543–550. ACM.

Kim, J. and André, E. (2008). Emotion recognition based on physiological changes in music listening. *TPAMI*, 30(12):2067–2083.

Kim, K. H., Bang, S. W., and Kim, S. R. (2004). Emotion recognition system using short-term monitoring of physiological signals. *Medical and biological engineering and computing*, 42(3):419–427.

Kishore, K. K. and Satish, P. K. (2013). Emotion recognition in speech using MFCC and wavelet features. In *IACC*, pages 842–847. IEEE.

Kobayashi, H. and Hara, F. (1992). Recognition of six basic facial expression and their strength by neural network. In *International workshop on robot and human communication*, pages 381–386. IEEE.

Kohler, C. G. et al. (2000). Emotion recognition deficit in schizophrenia: Association with symptomatology and cognition. *Biological psychiatry*, 48(2):127–136.

Kontopoulos, S. and Drakopoulos, G. (2014). A space efficient scheme for persistent graph representation. In *ICTAI*, pages 299–303. IEEE.

Kwon, O.-W., Chan, K., Hao, J., and Lee, T.-W. (2003). Emotion recognition by speech signals. In *Eighth European conference on speech communication and technology*.

Lane, R. D. et al. (1996). Impaired verbal and nonverbal emotion recognition in alexithymia. *Psychosomatic medicine*, 58(3):203–210.

Li, L., Zhao, Y., Jiang, D., Zhang, Y., Wang, F., Gonzalez, I., Valentin, E., and Sahli, H. (2013a). Hybrid deep neural network–Hidden Markov model (DNN-HMM) based speech emotion recognition. In *Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 312–317. IEEE.

Li, S., Yi, D., Lei, Z., and Liao, S. (2013b). The CASIA NIR-VIS 2.0 face database. In *CVPR*, pages 348–353.

Li, T. and Ogihara, M. (2004). Content-based music similarity search and emotion detection. In *ICASSP*, volume 5, pages V–705. IEEE.

Li, X. et al. (2007). Stress and emotion classification using jitter and shimmer features. In *ICASSP*, volume 4, pages IV–1081. IEEE.

Lin, Y.-L. and Wei, G. (2005). Speech emotion recognition based on HMM and SVM. In *International conference on machine learning and cybernetics*, volume 8, pages 4898–4901. IEEE.

Lin, Y.-P. et al. (2010). EEG-based emotion recognition in music listening. *Transactions on biomedical engineering*, 57(7):1798–1806.

Livingstone, S. R., Peck, K., and Russo, F. A. (2012). RAVDESS: The Ryerson audio-visual database of emotional speech and song. In *Annual meeting of the Canadian society for brain, behaviour, and cognitive science*, pages 205–211.

Martin, O., Kotsia, I., Macq, B., and Pitas, I. (2006). The eNTERFACE'05 audio-visual emotion database. In *ICDE*, pages 8–8. IEEE.

Mathe, E. and Spyrou, E. (2016). Connecting a consumer brain-computer interface to an internet-of-things ecosystem. In *PETRA*, pages 90–95. ACM.

Mohammadi, Z., Frounchi, J., and Amiri, M. (2017). Wavelet-based emotion recognition system using EEG signal. *Neural Computing and Applications*, 28(8):1985–1990.

Murugappan, M., Ramachandran, N., and Sazali, Y. (2010). Classification of human emotion from EEG using discrete wavelet transform. *Journal of biomedical science and engineering*, 3(04):390.

Nicholson, J., Takahashi, K., and Nakatsu, R. (2000). Emotion recognition in speech using neural networks. *Neural computing and applications*, 9(4):290–296.

Nwe, T. L., Foo, S. W., and De Silva, L. C. (2003). Speech emotion recognition using hidden Markov models. *Speech Communication*, 41(4):603–623.

Pan, Y., Shen, P., and Shen, L. (2012). Speech emotion recognition using support vector machine. *International Journal of Smart Home*, 6(2):101–108.

Petrushin, V. A. (2000). Emotion recognition in speech signal: Experimental study, development, and application. In *Sixth international conference on spoken language processing*.

Picard, R. W. (2003). Affective computing: Challenges. *International Journal of Human-Computer Studies*, 59(1-2):55–64.

Plutchik, R. (1980). A general psychoevolutionary theory of emotion. In *Theories of emotion*, pages 3–33. Elsevier.

Plutchik, R. (2001). The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 89(4):344–350.

Plutchik, R., Kellerman, H., and Conte, H. R. (1979). A structural theory of ego defenses and emotions. In *Emotions in personality and psychopathology*, pages 227–257. Springer.

Sato, N. and Obuchi, Y. (2007). Emotion recognition using mel-frequency cepstral coefficients. *Information and Media Technologies*, 2(3):835–848.

Schuller, B., Rigoll, G., and Lang, M. (2003). Hidden Markov model-based speech emotion recognition. In *ICASSP*, volume 2, pages II–1. IEEE.

Sprengelmeyer, R., Rausch, M., Eysel, U. T., and Przuntek, H. (1998). Neural structures associated with recognition of facial expressions of basic emotions. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 265(1409):1927–1931.

Spyrou, E. et al. (2018). A non-linguistic approach for human emotion recognition from speech. In *IISA*, pages 1–5. IEEE.

Stuhlsatz, A. et al. (2011). Deep neural networks for acoustic emotion recognition: Raising the benchmarks. In *ICASSP*, pages 5688–5691. IEEE.

Tao, J. and Tan, T. (2005). Affective computing: A review. In *International Conference on Affective computing and intelligent interaction*, pages 981–995. Springer.

Tian, C., Fan, G., Gao, X., and Tian, Q. (2012). Multiview face recognition: From Tensorface to v-Tensorface and k-Tensorface. *Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(2):320–333.

Vasilescu, M. A. O. and Terzopoulos, D. (2002). Multilinear analysis of image ensembles: Tensorfaces. In *European Conference on Computer Vision*, pages 447–460. Springer.

Wallbott, H. G. and Scherer, K. R. (1986). Cues and channels in emotion recognition. *Journal of personality and social psychology*, 51(4):690.

Wang, Y. and Guan, L. (2008). Recognizing human emotional state from audiovisual signals. *IEEE transactions on multimedia*, 10(5):936–946.

Wu, S., Falk, T. H., and Chan, W.-Y. (2011). Automatic speech emotion recognition using modulation spectral features. *Speech communication*, 53(5):768–785.

Yang, Y.-H., Lin, Y.-C., Su, Y.-F., and Chen, H. H. (2008). A regression approach to music emotion recognition. *Transactions on audio, speech, and language processing*, 16(2):448–457.

Zhalehpour, S., Onder, O., Akhtar, Z., and Erdem, C. E. (2016). BAUM-1: A spontaneous audio-visual face database of affective and mental states. *IEEE Transactions on Affective Computing*, 8(3):300–313.