Thesis

Keywords: Speech Emotion Recognition (SER), Emotional Speech, Emotion classification, Emotion Recognition Task (ERT), Language-Agnostic/Language-independent emotion

Prof guru: Paul Ekman, Plutchik's Wheel of Emotions, <https://scholar.google.it/citations?user=TxKNCSoAAAAJ&hl=it&oi=ao>, <https://scholar.google.it/citations?user=1-hcASEAAAAJ&hl=it&oi=ao>

Articoli fabio: <https://www.researchgate.net/publication/318031044_Online_Measuring_of_Available_Resources/figures?lo=1&utm_source=google&utm_medium=organic>

Libreria pirata dove cercare se non trovi: <http://185.39.10.101/>

App per ordinare paper: <https://paperpile.com/app>

codice sconto paperpile: Fabio\_Catania\_20

Workflow: nello State of Art metti una survey della parte sulla emotion classification, sugli speech corpora (per queste due parti puoi ricalcare il lavoro di Alberto), una parte su algoritmi e features (di solito trovi lo stato dell’arte negli articoli che approfondiscono una specifica parte), una parte sulle non-sentences, language agnostic emotions.   
Per farlo guarda prima di tutto qui: <https://dl.acm.org/journal/csur>, se non trovi guarda su google scholar (in particolare sono famose ACM digital library e IEEE). Generalmente cerca articoli molto recenti ( > 2015 circa) a meno che l’argomento non sia definito (ex: emotion classification)

Papers:

* R. R. Cornelius. The science of emotion: Research and tradition in the psy-chology of emotions. 1996.
* [(PDF) Describing the emotional states that are expressed in speech](https://www.researchgate.net/publication/222301235_Describing_the_emotional_states_that_are_expressed_in_speech)
* Douglas-Cowie E., Campbell N., Cowie R. & Roach P.(2003), Emotional speech: towards a new generationof databases, “Speech communication” 40, pp. 33-60
* Giovanni Costantini et al. “EMOVO Corpus: an Italian Emotional Speech Database”. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14). Reykjavik, Iceland: European Language Resources Association (ELRA), May 2014, pp. 3501–3504.
* Stanislavski’s system. url: https://en.wikipedia.org/wiki/Stanislavski% 5C%27s\_system. → PER IL TASK 1
* Paul Ekman and Wallace Friesen. “Unmasking the Face: A Guide to Recog- nizing Emotions From Facial Clues”. In: (Jan. 2003). → cosa di raccontare le storie prima di fare eseguire la performance (the stimulus must be natural)
* Antonio Origlia, Vincenzo Galat`a, and Bogdan Ludusan. “Automatic classi- fication of emotions via global and local prosodic features on a multilingual emotional database”. In: Jan. 2010. → PER IL TASK 4
* Paula M. Niedenthal. “Embodying Emotion”. In: Science 316.5827 (2007), pp. 1002–1005. issn: 0036-8075. doi: 10.1126/science.1136930. eprint: https : / / science . sciencemag . org / content / 316 / 5827 / 1002 . full . pdf. url: https://science.sciencemag.org/content/316/5827/1002. → Research on the embodiment of emotion has shown that subjects re- port that when they embody emotion through emotion-specific postures, they feel the associated emotion [32 ]

Talk about Ekman’s theory

<https://www.psychologytoday.com/us/blog/hide-and-seek/201601/what-are-basic-emotions>

<https://www.verywellmind.com/an-overview-of-the-types-of-emotions-4163976>

<https://www.verywellmind.com/theories-of-emotion-2795717>

LEGENDA:

* sottolineato in giallo = non capito
* sottolineato in rosso = reference da aggiungere
* barrato = probabilmente da rimuovere, ma da chiedere

2. State of the Art

2.1 Emotion Classification

Emotions play an incredibly powerful force on human behavior: strong emotions can cause you to take actions you might not normally perform or to avoid situations you enjoy. Emotion is often defined as a complex state of feelings that results in physical and psychological changes that influence thought and behavior. In this work the focus is, in particular, on the responses within the body that may be the cause or effect of emotions since this work is focused on speech emotion recognition (SER) that is a collection of methodologies that process and classify speech signals to detect the embedded emotions. ~~Emotionality is associated with a range of psychological phenomena, including temperament,~~ [~~personality~~](https://www.verywellmind.com/what-is-personality-2795416)~~, mood, and~~ [~~motivation~~](https://www.verywellmind.com/what-is-motivation-2795378)~~. According to author David G. Myers, human emotion involves "...physiological arousal, expressive behaviors, and conscious experience."~~  
For the purpose of this work, the first step consists then in identifying the most suited way to classify emotions according to what can be found in the literature. This paragraph analyzes the most common theories related to the classification of emotions.

2.1.1 Categorical models of emotion

The categorical model has the advantage that it represents human emotions automatically with easy to understand emotion labels. The emotional categories consist of distinct elements. The con is that not all the emotions are included as they are grouped together by single categories. So a categorical model has the limits (and advantages) of an identification task in attempting to distinguish the exact emotional states perceived by people.

Among these models, the one most commonly studied in literature is the one proposed by Paul Ekman. During the 1970s, psychologist Paul Eckman identified six basic emotions that he suggested were universally experienced in all human cultures: the idea was that all humans are thought to have an innate set of basic emotions that are cross-culturally recognizable. The emotions he identified were happiness, sadness, disgust, fear, surprise, and anger (Wallace V. Friesen Paul Ekman and Phoebe Ellsworth. “Emotion in the Human Face: Guidelines for Research and an Integration of Findings”. In: 1972.). This model is known as the discrete model of emotions. Later, he expanded the list to 15 emotions: amusement, anger, contempt, contentment, disgust, embarrassment, excitement, fear, guilt, pride in achievement, relief, sadness/distress, satisfaction, sensory pleasure and shame [. Ekman, P. Basic emotions. In Handbook of Cognition and Emotion; Dalgleish, T., Power, M., Eds.; John Wiley&Sons Ltd.: Hoboken, NJ, USA, 1999.]. So, the idea here was to think of emotions as discrete categories and in 1996, Cornelius, thought of a name for this set of basic emotions: the big six (R. R. Cornelius. The science of emotion: Research and tradition in the psychology of emotions. 1996).

Ekman’s theory on emotions is a milestone: most of the literature of the following years referred to his studies when taking into account these six emotions often adding a neutral state, particularly useful when signal analysis was done in order to identify a state of no emotion.

2.1.2 Dimensional models of emotion

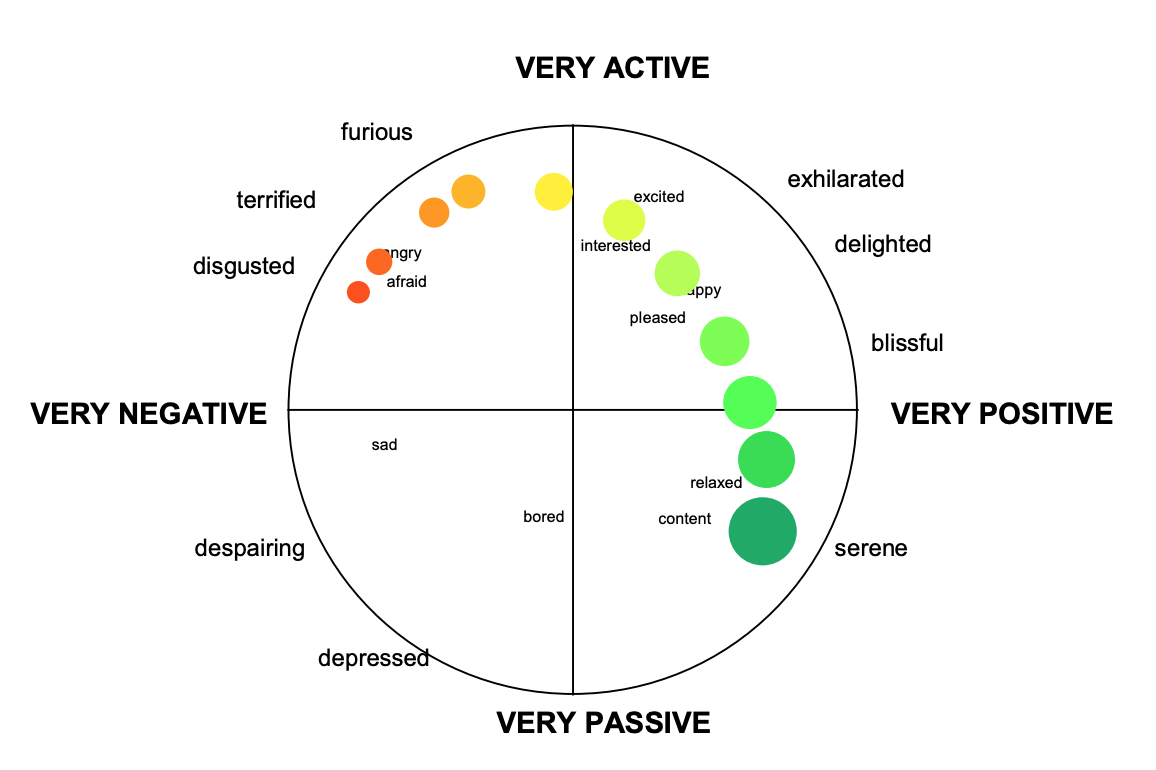
A second category of method used for recognizing emotions are the dimensional models: they define emotions according to one or more dimensions. This second type of approach appears to be more popular today [CIT: Gunes, H. and Schuller, B. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. Image and Vision Computing 31, 2 (2013), 120–136.] Several dimensional models of emotion have been developed, though there are just a few that remain as the dominant models currently accepted by most.[[9]](https://en.wikipedia.org/wiki/Emotion_classification#cite_note-Rubin_Telerico-9) The two-dimensional models that are most prominent are the circumplex model, the vector model, and the Positive Activation – Negative Activation (PANA) model.[[9]](https://en.wikipedia.org/wiki/Emotion_classification#cite_note-Rubin_Telerico-9)

The Circumplex model

The circumplex model of emotion was developed by James Russell (Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39*(6), 1161–1178. [https://doi.org/10.1037/h0077714](https://psycnet.apa.org/doi/10.1037/h0077714)). This model suggests that emotions are distributed in a two-dimensional circular space, containing arousal and valence dimensions. Arousal represents the vertical [axis](https://en.wikipedia.org/wiki/Cartesian_coordinate_system) and valence represents the horizontal axis, while the center of the circle represents a neutral valence and a medium level of arousal. In his work (Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39*(6), 1161–1178. [https://doi.org/10.1037/h0077714](https://psycnet.apa.org/doi/10.1037/h0077714)) Russell suggests that the interrelationships among affective dimensions can be represented by a spatial model in which affective concepts fall in a circle in the following order: pleasure (0), excitement (45), arousal (90), distress (135), displeasure (180), depression (225), sleepiness (270), and relaxation (315). In this model, emotional states can be represented at any level of valence (positive and negative) and arousal (high or low), or at a neutral level of one or both of these factors. Circumplex models have been used most commonly to test stimuli of emotion words, emotional facial expressions, and [affective](https://en.wikipedia.org/wiki/Affect_(psychology)) states.[[11]](https://en.wikipedia.org/wiki/Emotion_classification#cite_note-Remington-11)

To better visualize this, as pointed out in the description of the Belfast database (Ellen Douglas-Cowie, Roddy Cowie, and Marc Schro ̈der. “A New Emotion Database: Considerations, Sources and Scope”. In: Jan. 2000, pp. 39–44. ) a tool that can be used when dealing with dimensional model of emotion is FEELTRACE [Roddy Cowie et al. “’FEELTRACE’: An instrument for recording perceived emotion in real time”. In: Jan. 2000. ]. FEELTRACE (Figura) is a computer program based on this representation: the activation dimension measures how dynamic the emotional state is (the arousal); the evaluation dimension is a global measure of the positive or negative feeling associated with the state (the valence) and it allows users to generate time-varying descriptions of emotional content as they perceive it. Activation-evaluation space is represented by a circle on a computer screen, and observers (stimuli were 16 clips from TV programs) describe perceived emotional state by moving a pointer to the appropriate point in the circle using a mouse. The output records the position of the pointer on the two axes at intervals of a few milliseconds.

~~Russell and~~ [~~Lisa Feldman Barrett~~](https://en.wikipedia.org/wiki/Lisa_Feldman_Barrett) ~~describe their modified circumplex model as representative of core affect, or the most elementary feelings that are not necessarily directed toward anything. Different prototypical emotional episodes, or clear emotions that are evoked or directed by specific objects, can be plotted on the circumplex, according to their levels of arousal and pleasure.~~[~~[12]~~](https://en.wikipedia.org/wiki/Emotion_classification#cite_note-12)



Vector model

The vector model of emotion appeared in 1992.[[Bradley, M. M.; Greenwald, M. K.; Petry, M.C.; Lang, P. J. (1992). "Remembering pictures: Pleasure and arousal in memory". *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 18 (2): 379–390.](https://en.wikipedia.org/wiki/Emotion_classification#cite_note-Bradley-13) [doi](https://en.wikipedia.org/wiki/Doi_(identifier))[:](https://en.wikipedia.org/wiki/Emotion_classification#cite_note-Bradley-13)[10.1037/0278-7393.18.2.379](https://doi.org/10.1037%2F0278-7393.18.2.379)[.](https://en.wikipedia.org/wiki/Emotion_classification#cite_note-Bradley-13) [PMID](https://en.wikipedia.org/wiki/PMID_(identifier))[1532823](https://pubmed.ncbi.nlm.nih.gov/1532823)[.]](https://en.wikipedia.org/wiki/Emotion_classification#cite_note-Bradley-13) This two-dimensional model consists of [vectors](https://en.wikipedia.org/wiki/Vector_space) that point in two directions, representing a "boomerang" shape when drawing the possible combinations. The model assumes that there is always an underlying arousal dimension, and that valence determines the direction in which a particular emotion lies. A positive valence, for instance, would shift the emotion up the top vector and a negative valence would shift the emotion down the bottom vector [[ Rubin, D. C.; Talerico, J.M. (2009).](https://en.wikipedia.org/wiki/Emotion_classification#cite_note-Rubin_Telerico-9) ["A comparison of dimensional models of emotion"](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2784275)[. *Memory*. 17 (8): 802–808.](https://en.wikipedia.org/wiki/Emotion_classification#cite_note-Rubin_Telerico-9) [doi](https://en.wikipedia.org/wiki/Doi_(identifier))[:](https://en.wikipedia.org/wiki/Emotion_classification#cite_note-Rubin_Telerico-9)[10.1080/09658210903130764](https://doi.org/10.1080%2F09658210903130764)[.](https://en.wikipedia.org/wiki/Emotion_classification#cite_note-Rubin_Telerico-9) [PMC](https://en.wikipedia.org/wiki/PMC_(identifier))[2784275](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2784275)[.]](https://en.wikipedia.org/wiki/Emotion_classification#cite_note-Rubin_Telerico-9). In this model, high arousal states are differentiated in a neat way by their valence, while low arousal states are more neutral and are represented and concentrated near the meeting point of the vector. So, the vector model holds that there is an underlying dimension of arousal and a binary choice of valence that determines direction: this results in two vectors that both start at zero arousal and neutral valence and proceed as straight lines, one in a positive, and one in a negative valence direction creating the ‘boomberang’ shape. According to this, one main difference between the circumplex and vector model lies in the possibility to have or not emotions with high arousal and neutral valence: there are emotions such as aroused, astonished, and excited, that are emotionally intense yet neither very positive or negative: such points can be found in the circumplex model, but the vector model holds that at high arousal, positive and negative valences are distinct from one another and that true neutrality cannot be intensely felt. (TODO: MAGARI METTI UNA FIGURA CON IL BOOMERANG CHE FA CAPIRE CHE NON PUOI AVERE ALTA AROUSEL E VALENZA NEUTRA).

Positive activation – negative activation (PANA) model

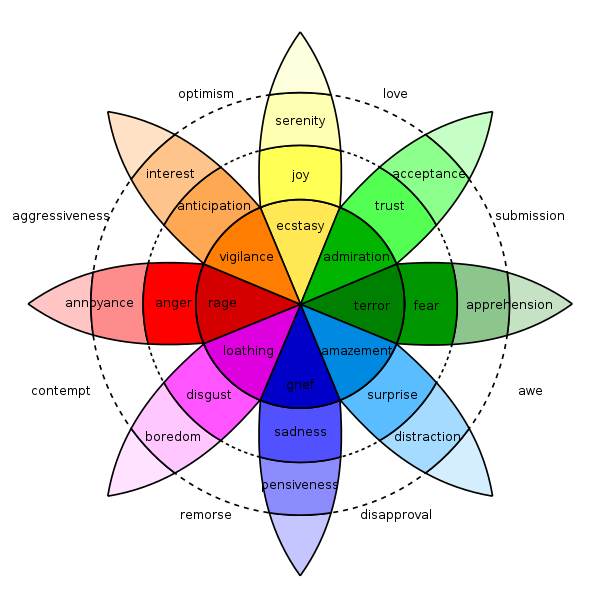
Wikipedia cit: The positive activation – negative activation (PANA) or "consensual" model of emotion, originally created by Watson and Tellegen in 1985,[[14]](https://en.wikipedia.org/wiki/Emotion_classification#cite_note-Watson-14) suggests that [positive affect](https://en.wikipedia.org/wiki/Positive_affectivity) and [negative affect](https://en.wikipedia.org/wiki/Negative_affectivity) are two separate systems. Similar to the vector model, states of higher arousal tend to be defined by their valence, and states of lower arousal tend to be more neutral in terms of valence.[[9]](https://en.wikipedia.org/wiki/Emotion_classification#cite_note-Rubin_Telerico-9) In the PANA model, the vertical axis represents low to high positive affect and the horizontal axis represents low to high negative affect. The dimensions of valence and arousal lay at a 45-degree rotation over these axes.[[14]](https://en.wikipedia.org/wiki/Emotion_classification#cite_note-Watson-14)

Guarda anche nel paper: A\_Comparison\_of\_Dimensional\_Models\_of\_Emotion\_Evid.pdf

Todo: non ho accesso al paper per cui non capisco

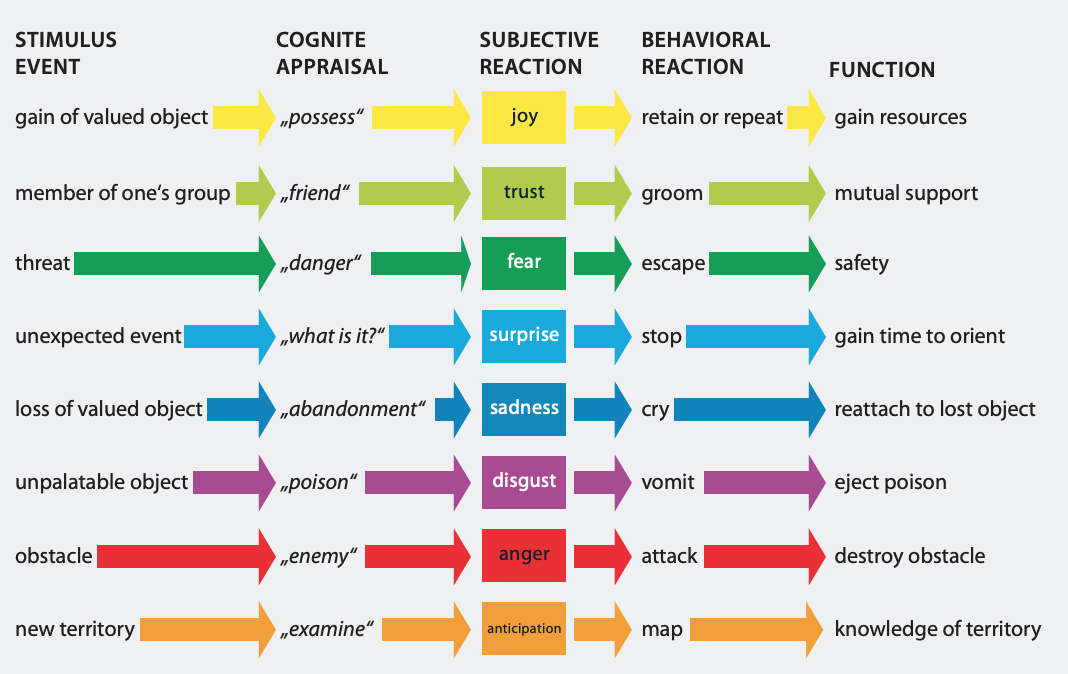
Plutchik's model

[Robert Plutchik](https://en.wikipedia.org/wiki/Robert_Plutchik) offers a three-dimensional model that is a hybrid of both basic-complex categories and dimensional theories. He started from the idea of emotion as a concept applicable to all evolutionary levels and to all animals as well as humans: emotions have an adaptive role in helping organisms deal with key survival issues posed by the environment, ‘organisms at all evolutionary levels face certain common functional survival problems’ [“EMOTION: A Psychoevolutionary Synthesis“ by Robert Plutchik; Harper & Row, Publishers (1980)]. Plutchik’s model arranges emotions in concentric circles where inner circles are more basic and outer circles more complex. It shows there are 8 basic emotions: joy, trust, fear, surprise, sadness, anticipation, anger, and disgust. The Plutchik’s wheel of emotions is represented in Figure xxx, but how can we interpret it?

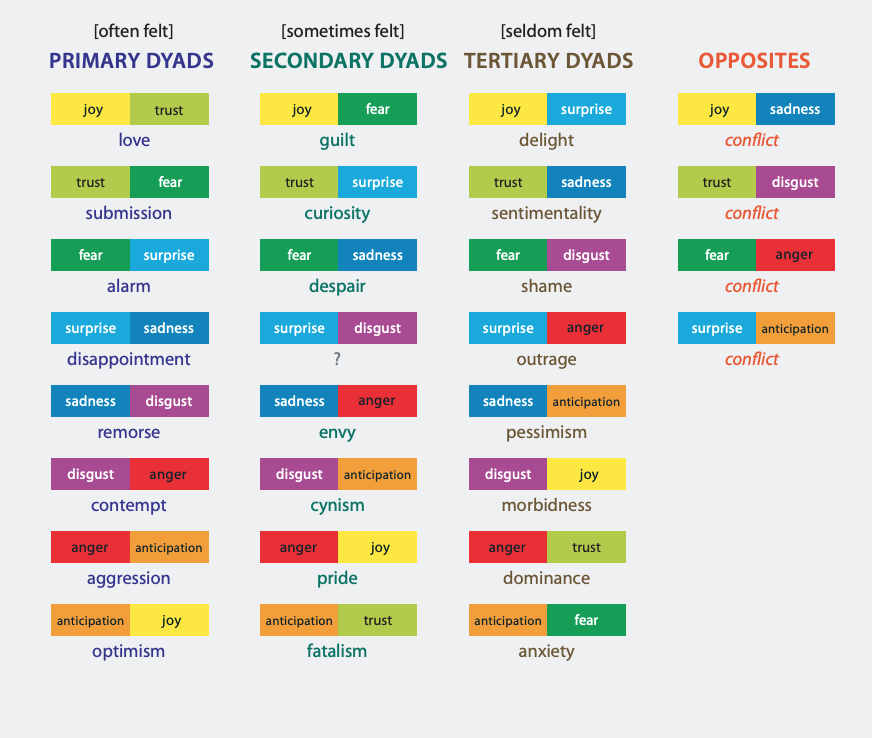


The eight sectors (middle layer) are designed to indicate that there are eight primary/basic emotions: anger, anticipation, joy, trust, fear, surprise, sadness and disgust. Each one of these basic emotions has a polar opposite. These are based on the physiological reaction each emotion creates in animals. So joy is the opposite of sadness, fear is the opposite of anger, anticipation is the opposite of surprise and finally disgust is the opposite of trust. For what concerns the emotions with no color, they represent an emotion that is a mix of two primary emotions: joy and trust combine, for instance, into love while sadness and disgust combine into remorse. Traversing the cones’ layers represents intensity: emotions intensify as they move from the outside layers to the center of the wheel. This is also designated by the color: the darker the shade, the more intense the emotion. For instance, fear at its lowest level of intensity is apprehension, while, at its highest level of intensity, it becomes terror.

According to the relation studied by Pluchick between emotions and the survival instincts of animals, he also proposed that eight defense mechanisms were manifestations of the eight core emotions as shown in the figure [["Robert Plutchik's Psychoevolutionary Theory of Basic Emotions"](http://www.adliterate.com/archives/Plutchik.emotion.theorie.POSTER.pdf) (PDF). *Adliterate.com*]



It is finally worth pointing out, that not all the possible mixtures of emotions (dyads) are represented in the fig xxx, an extensive list is shown instead in the following picture [["Robert Plutchik's Psychoevolutionary Theory of Basic Emotions"](http://www.adliterate.com/archives/Plutchik.emotion.theorie.POSTER.pdf) (PDF). *Adliterate.com*]



The PAD model

The [PAD emotional state model](https://en.wikipedia.org/wiki/PAD_emotional_state_model) is another dimensional model developed by [Albert Mehrabian](https://en.wikipedia.org/wiki/Albert_Mehrabian) and [James A. Russell](https://en.wikipedia.org/wiki/James_A._Russell) to describe and measure [emotional states](https://en.wikipedia.org/wiki/Affect_measures): they affirmed [Russell, James & Mehrabian, Albert. (1977). Evidence for a Three-Factor Theory of Emotions. Journal of Research in Personality. 11. 273-294. 10.1016/0092-6566(77)90037-X. ] with two studies that their work ‘*provided evidence that three independent and bipolar dimensions, pleasure-displeasure, degree of arousal, and dominance-submissiveness, are both necessary and sufficient to adequately define emotional states*’.

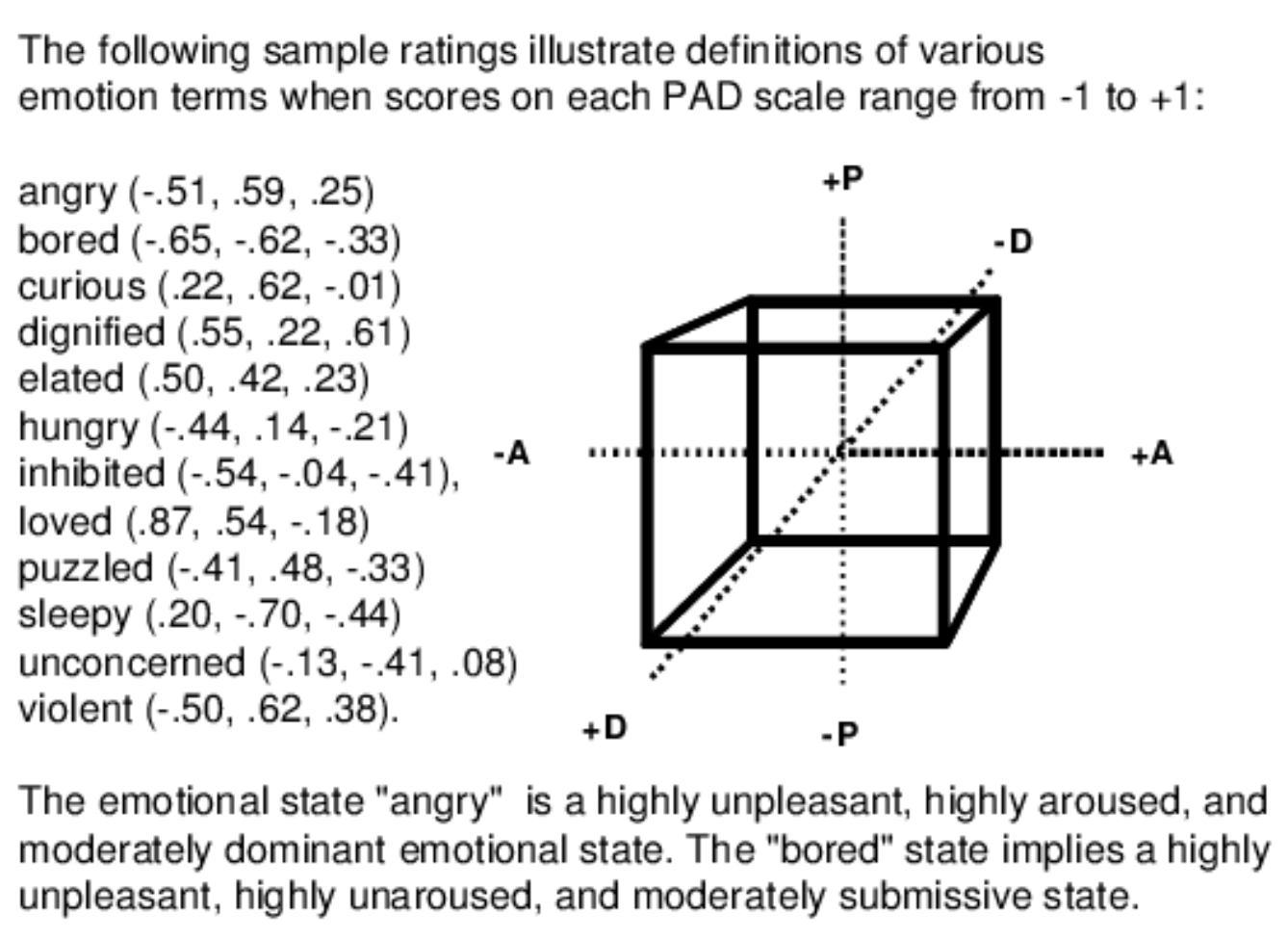
The three dimensions are defined as bipolar such that each of them is a continuum.

Pleasure ranges from extreme pain or unhappiness at one end to extreme happiness or ecstasy at the other end. So, the Pleasure scale measures how pleasant an emotion may be. To better understand, both anger and fear, for instance, are unpleasant emotions and score high on the displeasure scale (the negative part of the pleasure scale). Joy on the opposite is a pleasant emotion.

The Arousal scale ranges from sleep through intermediate states of drowsiness and then alertness to frenetic excitement at the opposite pole. The Arousal-Nonarousal scale measures then the intensity of the emotion: for instance while both anger and rage are unpleasant emotions, rage has a higher intensity or a higher arousal state. However boredom, which is also an unpleasant state, has a low arousal value.

Finally we have Dominance that ranges from feelings of total lack of control or influence on events and surroundings to the opposite extreme of feeling influential and in control. The Dominance-Submissiveness scale represents the controlling and dominant nature of the emotion: for instance while both fear and anger are unpleasant emotions, anger is a dominant emotion, while fear is a submissive emotion.

A representation of this model with some sample ratings can be seen in figure xxx



It must be pointed out that these three values are completely independent and any value on one dimension may occur simultaneously with any value on either of the other two dimensions.

2.2 Cultural and Language Agnostic Emotions

Psychologists have long debated whether emotions are universal versus whether they vary by culture. Another related field of research concerns the possibility of building a speech emotion recognizer that performs well independently from the language used: a language agnostic speech emotion recognizer. In this work I just want to focus on the idea that there may be some emotions that are universal, that are cross-cultural and common to all the human beings. Assuming that this is true, my purpose is to assess the performances of the models built in this project to verify if it may be possible to build a recognizer that is universally good independently from the language of the speaker (and consequently from his/her cultural background): this will be done by testing the model on non-sentences and non-words instead of using different languages. In this way it will be possible to investigate the importance that semantics plays in the SER task (non-sentences) and give at least a clue about the importance that the language has in this task (non-words). A clue about the possibility of obtaining a successful model in this sense can be found in [Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations Disa A. Sauter, Frank Eisner, Paul Ekman, Sophie K. Scott Proceedings of the National Academy of Sciences Feb 2010, 107 (6) 2408-2412; DOI: 10.1073/pnas.0908239106] where, through an experiment, the authors assessed the recognition of nonverbal emotional vocalizations, such as screams and laughs, across two completely different cultural groups: western participants were compared to individuals from remote, culturally isolated Namibian villages. This work doesn’t verify the recognition of non-sentences and non-words, but it is something really close to it: it focuses on the ‘sounds’ of emotions, so something separated from the specific language and semantic, and showed that vocalizations communicating the so-called “basic emotions” (anger, disgust, fear, joy, sadness, and surprise) were bidirectionally recognized by both groups. So, they showed that this set of emotions were cross-culturally recognized from the vocal signals that were perceived as communicating specific affective states. So, “this finding supports theories proposing that these emotions are psychological universals and constitute a set of basic, evolved functions that are shared by all humans” and, in particular, it also gives a hint, useful for my research, to the study of the recognition of these universal emotions from sound samples that are completely unrelated to language and semantic since the authors used non-verbal vocalizations in their experiments.

TODO: dire di più su questa ricerca sulle nonverbal emotional vocalizations???  
 ~~This idea will be applied to the SER field testing it in some way with the models developed: samples of non-sentences and non-words will be tagged and used to train the models and its performances will be assessed to give a clue about the importance or non-relevance of the semantic meaning of the speech samples. This will give a hint about the importance that culture (and consequenthe language and the sounds of voice in this case) have in the emotion recognition task.~~

2.2.1 Are There Universal Emotions?

There is a long debate in the literature whether emotions are universal versus whether they vary by culture. In [The Expression of the Emotions in Man and Animals](https://en.wikipedia.org/wiki/The_Expression_of_the_Emotions_in_Man_and_Animals) published in 1872, [Charles Darwin](https://en.wikipedia.org/wiki/Charles_Darwin) theorized that emotions were evolved traits universal to the human species: he proposed universal facial expressions of emotion on the basis of his evolutionary theory. Paul Ekman, in his ‘Universal Facial Expressions of Emotion’ [CIT] affirmed that there are ‘evoking stimuli’ elicitors for emotions (in particular in the facial expressions related to emotions) that are universally associated with particular emotions, while there are also some of these ‘stimuli’ that are learned and so that vary through different cultures. Ekman, through a series of experiments, concluded that: “ *there are distinctive movements of the facial muscles for each of a number of primary states, and these are universal to mankind*”. So he claimed that the facial patterns elicited by the set of basic emotions (happiness, sadness, anger, fear, surprise and disgust) are pancultural.

On the other hand, a lot of the literature refers to theories according to which [facial expressions](https://en.wikipedia.org/wiki/Facial_expressions) and their meanings were determined through behavioral learning processes. A prominent advocate of the latter perspective was the anthropologist [Margaret Mead](https://en.wikipedia.org/wiki/Margaret_Mead), who had travelled to different countries examining how cultures communicated using nonverbal behavior. She also wrote against Ekman for disagreeing with [Ray Birdwhistell](https://en.wikipedia.org/wiki/Ray_Birdwhistell)'s theory: Birdwhistell was an anthropologist and was one of the most influential writer arguing for the culture specific view of emotion’s facial expressions and the non existence of universal symbols of emotional state [CIT: Birdwhistell, Ray L. The kinesic level in the investigation of the emotions. Chapter 7, Part II in *Expression of the Emotions in Man,* Peter H. Knapp, MD, (Ed.), New York: International Universities Press, Inc., 1963 → TODO: CIT COPIATA; non ho trovato l’originale].

A lot of research in this field focused on the basic assumption that emotions are universally recognized from facial expressions, but the question may be way more complicated than this as stated by Russell [Russell JA. Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. Psychol Bull. 1994 Jan;115(1):102-41. doi: 10.1037/0033-2909.115.1.102. PMID: 8202574.]. In this review, Russell described the key evidence on which the universality theory rests and raised questions about the methods used to gather that evidence. He first concluded that “*this is a topic on which opinions can differ*” and, even if many may find the universality thesis the most plausible alternative available, he stated that “*we need to abandon any implicit assumption that we have only two alternatives: randomness and universality*” by proposing eight alternatives options. So, he basically put the emphasis on alternative broader and smoother conceptualizations that may be different according to the context etc. He proposed to overcome the influence of the implicit assumptions of universality and he said that “*the most interesting means to this end is to take seriously the conceptualizations (ethnotheories, cultural models) found in other cultures. Rather than ask whether a given culture agrees with one preformulated hypothesis, we might more usefully ask how members of that culture conceptualize emotions and facial behavior*.” because “ *Their theories are not to be believed any more than our own*.”.

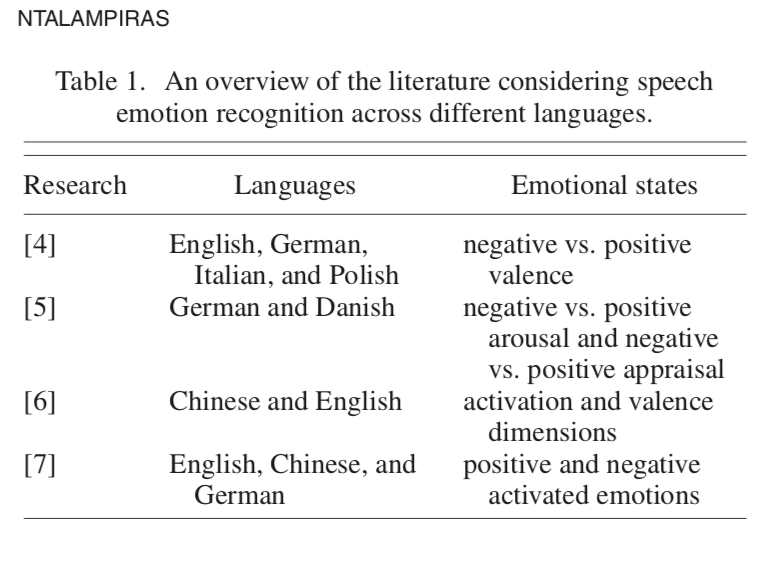
In conclusion, even if the universality thesis is the most popular, it is evident that there is not a total agreement about this topic in the literature. A lot of studies investigated this problem and subjects related or derived from this: the language agnostic SER research field is one of these.

2.2.2 Language Agnostic SER

Cross-language speech emotion recognition is receiving increased interest because of its extensive real-world applicability. As it is also

highlighted by one of the pioneers in speech emotion recognition (SER), Björn W. Schuller, one of the current challenges of research is vocal emotion analysis across cultures and languages [Björn W. Schuller. 2018. Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends. Commun. ACM 61, 5 (May 2018), 90–99. DOI:<https://doi.org/10.1145/3129340>]. In this work, Schuller states: “ several further steps must be taken before SER can be considered ready for broad consumer usage “in the wild.” These include robustness across cultures and languages as one of the major white spots in the literature. A number of studies show the downgrades one may expect when going cross-language in terms of acoustic emotion recognition. As to cross-cultural studies, these are still particularly sparse, and there exists practically no engine that is adaptive to cultural differences at the time”.  
The field of SER exploits signal processing and pattern recognition algorithms to predict the emotions conveyed by the speakers’ samples. Because of the increased attention in the last decades, several systems have been developed, achieving encouraging performance. Nonetheless a lot of these models and algorithms are based on homogenous speech corpora with speech samples of the same language: so, when thinking, to some of the real-world applications, such algorithms need to face data characterized by great diversity, including different languages. Some researches showed that good results may be obtained in this sense. As pointed out in

[Ntalampiras, Stavros. (2020). Toward Language-Agnostic Speech Emotion Recognition. Journal of the Audio Engineering Society. 68. 7-13. 10.17743/jaes.2019.0045.] that built a novel framework realizing language-agnostic SER and demonstrated the feasibility of recognizing emotions from speech samples in a language-independent setting through a combination of mel-scaled and modulation spectrograms feeding a GMM classification scheme, there are some approaches that address the cross-corpus problem (this referes to the use of databases representative of emotional states across different languages), a summary table of these is shown here (Table xxx).



PEZZO COPIATO (SE LO VUOI LASCIARE MODIFICALO)

Kernel canonical correlation analysis is explored in [4] to classify four emotional speech corpora with different languages (English, German, Italian, and Polish). They used 384 features, as in the Interspeech 2009 Emotion Chal- lenge using openSMILE [24], in combination with a Simple Logistic classifier. A cross-corpus classification of speech emotion is discussed in [5] on two emotional databases encompassing German and Danish languages. The authors used 266 frequency domain features and Sequential Mini- mal Optimization (SMO) , Multilayer Perceptron, Rotation Forest, FT tree, and Random Forest classifiers. A system in- tegrating multiple emotion perspectives of Chinese and En- glish is presented in [6]. Their feature set included MFCCs, pitch, and intensity, whereas a linear support vector regres- sion kernel achieved activation and valence prediction. The work explained in [7] employs data including three lan- guages (English, Chinese, and German). The extracted fea- tures are 38 low-level time and spectral descriptors as well as 21 functionals. The corresponding classifier is an SVM, and they focus on the similarities as well as differences between human perception and automatic classification.

FINE PEZZO COPIATO

TODO: puoi approfondire la ricerca già esistente in questo senso (guarda introduzione del paper ‘Toward Language-Agnostic Speech Emotion Recognition’). Oppure cerca altri paper magari

2.3 Emotional Speech Databases

In this section, exploiting the thorough surveys of [CIT Ayadi, Moataz M. H. El et al. “Survey on speech emotion recognition: Features, classification schemes, and databases.” Pattern Recognit. 44 (2011): 572-587 + Mehmet Berkehan Akçay, Kaya Oğuz,

Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers,

Speech Communication,

Volume 116,

2020,

Pages 56-76,

ISSN 0167-6393,

https://doi.org/10.1016/j.specom.2019.12.001.] I just provide a quick overview of the most important speech corpora available and of the design choices that should guide the implementation of this kind of databases.

Databases are an essential part of speech emotion recognition since classification process relies on the labeled data.

An important role in the evaluation of an emotional speech recognizer is the degree of naturalness of the database used to assess its performance. Incorrect conclusions may be established if a low-quality database is used. Moreover, the design of the database is critically important to the classification task being considered: it is strictly related to the task that we want to perform, the gathering of the data will differ according to which is the purpose of the recognizer and to the number and type of emotions included in the database.

The most relevant factors [CIT 69,22] to properly design a database of this kind are:

* **Natural emotions or acted ones?** It is more realistic to use speech data that are collected from real life situations because they provide very natural conveyed emotion. Nonetheless, there may be some legal and moral issues that prohibit their use for research purposes. Alternatively, emotional sentences can be elicited in sound laboratories with professional and semi-professional actors as in the majority of the existing databases. This solution is clearly more ‘simple’ even if we may criticize that acted emotions are not the same as real ones. Williams and Stevens [135] found that acted emotions tend to be more exaggerated than real ones. Anyway, the relationship between the acoustic correlate and the acted emotions does not contradict that between acoustic correlates and real ones. Besides that, this solution is far less complicated and the price to pay is not too high. Finally, a somehow hybrid solution consists instead in using elicited speech databases: these are created by placing speakers in a simulated emotional situation that can stimulate various emotions. In this way, even if the emotions are not fully-elicited, they are close to real ones.
* **Who utters the emotions?**: In most emotional speech databases, professional actors are invited to express pre-determined sentences with the required emotions. However, in some of them such as the Danish Emotional Speech (DES) database [38], semi- professional actors are employed instead in order to avoid exaggeration in expressing emotions and to partially be closer to real world situations. An alternative solution is gathering data from common people through crowdsourcing: good results were obtained through this approach [CIT Emozionalmente] and I’ll try to further investigate this approach in this research.
* **How to simulate the utterances?** The recorded utterances in most emotional speech databases are not produced in a conversational context [69]. So the speech samples may lack some naturalness since it is believed that most emotions are outcomes of our response to different situations and, as stated by Ekman [CIT Paul Ekman and Wallace Friesen. “Unmasking the Face: A Guide to Recog- nizing Emotions From Facial Clues”. In: (Jan. 2003)], ‘an emotion can be expressed genuinely only if the stimuls is natural’. Generally, there is the need to evoke in the speaker the emotion in some way. A possible solution, especially for experienced speakers, is to act as if they were in a specific emotional state by self-inducing the emotion. If professional actors are not available anyway the speaker may be invited to utter the emotional utterances in the same way. There are also other possible approaches used in order to help the actor reach the required emotional states. In [59] for instance, it was proposed to use computer games to induce natural emotional speech. Voice samples were elicited following game events whether the player won or lost the game and were accompanied by either pleasant or unpleasant sounds. Another example can be found in the creation of the eNTERFACE [O. Martin et al. “The eNTERFACE05 Audio-Visual Emotion Database”. In: Feb. 2006, pp. 8–8. isbn: 0-7695-2571-7. doi: 10.1109/ICDEW.2006.145.] database where the subjects of the experiments were asked to listen to six different stories (each of which with the purpose of evoking a different emotion) before starting to record their samples. A final example is the Belfast database where one of the sources of data were recorded interviews that were guided by a moderator whose intention was to convey in the listener stronger feelings and emotions. On the other hand, in many of the other research, a well-known technique was used: the Stanislavski Method [CIT? Stanislavski’s system. url: https://en.wikipedia.org/wiki/Stanislavski% 5C%27s\_system.]. The creator of the technique was an important Russian theatre actor and the goal of his method is to work on the actor's conscious thought to activate sympathetically and indirectly the other less-controllable psychological processes, such as emotional experience and subconscious behaviour.
* **Balanced utterances or unbalanced utterances?** While balanced utterances are useful for controlled scientific analysis and experiments, they may reduce the validity of the data. As an alternative, a large set of unbalanced and valid utterances may be used. (TODO non ho capito bene).
* **Utterances are uniformly distributed over emotions?** Some corpus developers (such as in the Berlin corpus [18]) prefer that the number of samples for each emotion is almost the same in order to fairly evaluate the classification accuracy. On the other hand, many other researchers prefer that the distribution of the emotions in the database reflects their frequency in the world [140,91]. For example, the neutral emotion is the most frequent emotion in our daily life. Hence, the number of data with neutral emotion should be the largest in the emotional speech corpus: in this way the trained classifier may have more variance when recognizing less common emotions, but it may boost its performances on data representing more frequent emotions.
* **Same statement with different emotions?**: In order to study the explicit effect of emotions on the acoustic features of the speech samples, it is common in many databases to record the same sentence with different emotions. One upside of this approach is to ensure that the human judgment on the perceived emotion is purely based on the emotional content of the sentence and not on its lexical content. Another solution to perceive this consists in using non-sentences or even non-words in the speech data in order to separate the emotion from its lexical content

In addition to these, other design issues may be considered such as the distribution of age and gender in the speakers. Also, most databases include adult speakers, but databases of children and elders exist as well (KISMET and BabyEars for instance considered infant-directed emotions).

In the following I reported a table summarizing the information related to some of the most relevant speech corpora available in the literature. Most of these emotional speech databases are private and not available for public use: this results in the lack of coordination among researchers in this field and is certainly no good.

NB: when the source is Actors it means that the emotions are acted in lab

TODO:~~COPIA TUTTI QUELLI DALLA SURVEY UTILIZZATA (pari pari, potresti anche usare quella più recente intitolata Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers)~~ Alla fine ho messo solo alcuni interessanti più quelli del paper che contenevano big six + neutral oppure un subset di queste emozioni

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Corpus | Access | Language | Size | Source | Emotions |
| AFEW Database | Free | ENG | 330 speakers, 1426 utterances | Natural: movies and TV shows | Anger, Disgust, Fear, Happiness, Neutral, Sadness Surprise, |
| Berlin Emotional Database (EmoDB) | Open Access | GER | 7 Emotions x 10 speakers(5 male, 5 female) x 10 utterances | Acted: Actors | Anger, Boredom, Disgust, Fear, Happiness, Sadness, Neutral |
| Belfast Database | - | ENG | 100 speakers, 239 utterances | Natural and Acted: Television Programs + Postgraduate Students | Dimensional = Activation- Evaluation space Categorical = list of 16-24 emotion labels |
| Chinese Annotated Spontaneous Speech corpus (CASS) | Commercially available | MANDARIN | 7 speakers (2 male, 5 female),6 h of speech | Natural: recordings of university lectures by professors and invited speakers, student colloquia, and other public meetings | Anger, Fear, Happiness, Neutral, Sadness, Surprise, |
| Chinese Elderly Emotional Speech Database (EESDB) | Free to research use | MANDARIN | 16 speakers (8 male, 8 female),400 utterances from teleplay | Natural: chinese TV’s statements | Anger, disgust,fear, happiness, neutral, sadness, surprise |
| Chinese Emotional Speech Corpus (CASIA) | Commercially available | MANDARIN | 6 Emotions x 4 Speakers (2 male, 2 female) x 500 utterances | Acted: Actors | Surprise, happiness,sadness, anger,fear, neutral |
| Danish Emotional SpeechDatabase (DES) | Free | DANISH | 4 speakers (2 male, 2 female) 10 min of speech | Acted: Actors | Anger, Happiness, Neutral, Sadness, Surprise |
| Electromagnetic Articulography Database (EMA) | Free to research use | ENG | 3 speakers (1 male, 2 female) 14 sentences for male, 10 sentences for female | Acted: Actors | Anger, Happiness, Neutral, Sadness |
| eNTERFACE | Public and free | ENG | 42 Speakers (14 different nationalities), 1166 utterances | Acted: participants of a workshop on multimodal interfaces, mostly research engineers | Anger, Disgust, Fear, Joy, Sadness, Surprise |
| EMOVO | Public and free | ITA | 6 Actors \* 14 Sentences \* 7 emotions = 588 utterances | Acted: Actors | Anger, Disgust, Fear, Joy, Sadness, Surprise + Neutral |
| €motion | - | ITA, ENG, FR, GER | 1560 total utterances | Acted: Actors + Naive Speakers | Anger, Disgust, Fear, Happiness, Sadness, Surprise |

TODO: forse devi scrivere cose sulla validation?

From this table, we can see that the emotions are usually simulated by professional or nonprofessional actors. In fact, as pointed before, there are some legal and ethical issues that may prevent researchers from recording data coming from the ‘real’ world. In addition, nonprofessional actors are invited to produce emotions in many databases in order to avoid exaggeration in the perceived emotions. We notice also that most of the databases share the basic emotions as described in Section one with some variations, in the Belfast database for instance the authors also used a dimensional model for classifying emotions in order to give relevance to the cases where emotions are not archetypal, but are more controlled, weaker and subtler because, as the authors pointed out, “displays of intense emotion are rare“ and “clear examples of ‘pure’ primary emotions are even rarer”. Finally it is worth pointing out that all the databases are enclosed in the adult-directed emotions framework as the one described in Section one (or similar).

According to the surveys of [cit 1][cit 2] there are some common limitations to these speech corpora. Besides the already mentioned problem of lack of public availability of usage that creates a lot of coordination problems,   
most speech emotional databases do not simulate well enough emotions in a natural and clear way. This is evident by the relatively low recognition rates of human subjects and it is due also to the fact that emotions may be ambiguous by nature: this is actually an intrinsic problem of the SER task.   
~~Moreover, in some databases such as KISMET, the quality of the recorded data is not so good. Finally, it happens that phonetic transcriptions are not provided with some databases such as BabyEars: it is difficult to extract linguistic content that may be useful for analysis from the samples of such databases.~~

2.4 Emotion Recognition Task: Features And Models

We define a SER system as a collection of methodologies that process and classify speech signals to detect emotions embedded in them. This process involves various steps. After the creation of a database as we saw in the last section, there are basically three more steps: data preprocessing, features extraction and the classification algorithm.

TODO: ~~Riparafrasa tutto perchè copiato per ora~~ Fatto, ma fai un secondo giro

2.4.1 Data Preprocessing

Preprocessing is the first step that should be performed after collecting data that will be used to train the classifier in a SER system. Some of these preprocessing techniques are used for feature extraction, while others are used to normalize the features so that variations of speakers and recordings would not affect the recognition process.

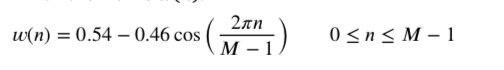
**Framing**

Signal framing, also known as speech segmentation, is the process of dividing speech signals that are continuous into fixed length segments to overcome several challenges in SER.  
Emotion can change in the course of speech since the signals are non-stationary. However, speech remains invariant for a sufficiently short period (20 to 30 ms). By framing the speech signal, this quasi-stationary state can be approximated, and local features can be obtained related to the specific frame. *Additionally, the relation and information between the frames can be retained by deliberately overlapping 30% to 50% of these segments. Continuous speech signals restrain the usage of processing techniques such as Discrete Fourier Transform (DFT) for feature extraction in applications such as SER. Consequently, fixed size frames are suitable for classifiers, such as Artificial Neural Networks, while retaining the emo- tion information in speech.*

**Windowing**

After framing the speech signal, the next step generally consists in applying a window function to frames. The windowing function is used to reduce the effects of leakages that occur during Fast Fourier Transform (FFT) of data caused by discontinuities at the edge of the signals: the borders of segments are visible as discontinuities, which are incongruent with the real-world signal. To reduce the impact of segmenting on the statistical properties of the signal, windowing is applied to the temporal segments. Windowing functions are smooth functions which go to zero at the borders. By multiplying the input signal with a window function, the windowing function also goes to zero at the border such that the discontinuity at the border becomes invisible. Windowing thus changes the signal, but the change is designed such that its effect on the signal statistics is minimized.

Typically a Hamming window is used where the window size is *M* for the frame *w*(*n*).



**Voice activity detection**

*Voice activity detection* (VAD) refers to the task of determining whether a signal contains speech or not. A speech sample consists of three parts: voiced speech, unvoiced speech, and silence. Voiced speech is generated with the vibration of vocal folds that creates periodic excitation to the vocal tract during the pronunciation of phonemes which are perceptually distinct units of sound that distinguish one word from another; such as bag, tag, tab. On the other hand, unvoiced speech is the result of air passing through a constriction in the vocal tract, producing transient and turbulent noises that are aperiodic excitations of the vocal tract. Due to its periodic nature, voiced speech can be identified and extracted. The detection of the presence of voiced speech among various unvoiced speech and silence is called endpoint detection, speech detection or voice activity detection.

The performance of the endpoint detection algorithm affects the accuracy of the system. It’s hard to model silence and noise accurately in a dynamic environment; if voice and noise frames are removed, it will be easier to model speech → credo che l’idea sia rimuovere quelle parti di segnale che contengono rumori diversi dallo speech.. In addition, speech consists of many silent and noisy frames which increase the computational complexity. Removal of these frames decreases the complexity and increases accuracy. ~~Most widely used methods for voice activity detection are zero crossing rate, short time energy, and auto-correlation method.~~

~~Zero crossing rate is the rate at which a signal changes its sign from positive to negative or vice versa within a given time frame. In voiced speech, the zero crossing count is low whereas it has a high count in unvoiced speech (Bachu et al., 2010). The voiced speech has high en- ergy due to its periodicity while low energy is observed in the unvoiced speech. The auto-correlation method provides a measure of similarity between a signal and itself as a function of delay. It is used to find re- peating patterns. Because of its periodic nature, voiced signals can be detected using the auto-correlation method.~~

**Normalization**

Data normalization is an important step which is used to reduce speaker and recording variability without losing the discriminative strength of the features. By using feature normalization, the generalization capability of features is increased. This step is typically performed in every machine learning task. Normalization can be done at different levels, such as function level and corpus level. Most widely used normalization method is z-normalization (standard score). If the mean of the data is *𝜇* and the standard deviation is *𝜎*, z-normalization is calculated as

z = x - *𝜇 /*  *𝜎*

***Noise Reduction***

When recording an utterance the noise present in the environment is captured along with the speech signal. This affects the recognition rate, hence some noise reduction techniques must be used to eliminate or reduce the noise. Minimum mean square error (MMSE) and log-spectral amplitude MMSE (LogMMSE) estimators are most successfully applied methods for noise reduction (Pohjalainen et al., 2016).

~~In MMSE, the clean signal is estimated from a given sample function of the noisy signal. It needs a priori information of speech and noise spectrum. It is based on the assumption that the additive noise spectrum and estimate of the speech spectrum is available. The aim of the method is minimizing the expected distortion measure between clean and estimated speech signal.~~

~~There are also single-channel noise reduction techniques such as spectral subtraction that can be used for noise reduction.~~

**Feature Selection and dimension reduction**

Feature selection and dimension reduction are important steps in emotion recognition. There is a need to use a feature selection algorithm because there are many features and there is no certain set of features to model the emotions. Otherwise, with so many features, the classifiers are faced with the curse of dimensionality: the number of samples needed increases incrementally as the number of dimensions increases. Also, increased training time and overfitting may highly affect the prediction rate. Feature selection is the process of choosing a relevant and useful subset of the given set of features. The unneeded, redundant or irrelevant attributes are identified and removed to provide a more accurate predictive model: to reduce the variance of the model (i.e. its sensitivity to the change in the data) and increase its generalization capabilities.

2.4.2 Features

TODO: rimuovi esempi e lascia tutto riassunto nella tabellona

Features are an important aspect of speech emotion recognition. Carefully chosen sets of features help in successfully characterizing each emotion and increases the recognition rate. Various features have been used for SER systems; however, there is no generally accepted set of features for precise and distinctive classification. Speech is a continuous signal of varying length that carries both information and the emotional state of the speaker. Therefore, global or local features can be extracted depending on the required approach. Features are extracted from the speech signal in two forms:

1. **Local** features by partitioning the signals into smaller frames and computing statistics of each frame
2. **Global** features by calculating statistics on the whole utterance.

So **local** features, also known as short-term or segmental features, represent the temporal dynamics, where the purpose is to approximate a stationary state. On the other hand, **global** features, also called long-term or supra-segmental features, represent the gross statistics such as mean, minimum and maximum values, and standard deviation. The stationary states are important because emotional features are not uniformly distributed over all positions of the speech signal (Rao et al., 2013). For example, emotions such as anger are predominant at the beginning of utterances, while surprise is conveyed at the end of it. Hence, to capture the temporal information from the speech, local features are used.

Moreover, the soundwave’s features are also classified in two other distinctive classes:

1. Low- Level-Descriptors (LLDs): LLDs contain prosodic features, which are **global** and spectral features and their derivatives that are **local**.
2. Functionals: include statistical features that derive from LLDs and therefore, they are **global** features.

These local and global features of SER systems can be also divided in the following four categories.

* Prosodic Features
* Spectral Features
* Voice Quality Features
* Teager Energy Operator (TEO) Based Features

Prosodic and spectral features are used more commonly in SER systems. TEO features are usually employed to recognize stress and anger. Anyway, in practice these features are usually combined in a SER system.

**Prosodic Features**

Prosodic features are those that can be perceived by humans, such as intonation and rhythm. A typical example is raising the intonation in a sentence that is meant as a question.They are also called para-linguistic features as they deal with the elements of speech that are properties of large units as syllables, words, phrases, and sentences. ~~Since they are extracted from these large units, they are long-term features~~. Prosodic features have been discovered to convey the most distinctive properties of emotional content for speech emotion recognition   
The most widely used prosodic features are based on fundamental frequency, energy, and duration.

The **fundamental frequency F0** is defined as the number of times a sound wave produced by the vocal cords repeats during a given time period. F0 is created by the vibrations in the vocal cord: it is the central tendency of the frequency of vibration of the vocal folds during connected speech and it has a correlation with the perceived pitch of a speaker's voice. It yields rhythmic and tonal characteristics of the speech. The change of the fundamental frequency over the course of an utterance yields its fundamental frequency contour (a graph of fundamental frequency represented as time varies) whose statistical properties can be used as features.  
The **energy** of the speech signal, sometimes referred to as volume or intensity, provides a representation which reflects amplitude variation of the speech sound wave over time. Researchers suggest that high arousal emotions such as anger happiness or surprise yields increased energy while disgust and sadness result with decreased energy (Lin et al., 2012).  
**Duration** is the amount of time used to build vowels, words and similar constructs that are present in speech. Speech rate, duration of silence regions, rate of duration of voiced and unvoiced regions, duration of longest voiced speech are among the most widely used duration related features.

It is intuitive that there are correlations between prosodic features and emotional states. Prosodic features expose the changes during the course of emotional speech. For instance, throughout the production of highly active emotions such as anger, fear, anxiety, and joy, mean *F*0, *F*0 variability, and vocal intensity increases. *F*0 contour decreases over time during the expression of anger. In contrast, it increases over time during the expression of joy. On the other hand low-level arousal (passive) emotions such as sadness yields lower mean *F*0, *F*0 variability, and vocal intensity compared to natural speech, while also *F*0 decreases over time (Frick, 1985; Bachorowski, 1999). Duration to express anger is shorter than duration to express sadness (Rao et al., 2013).

Busso et al. analyzed various expressive *F*0 contour statistics to find the emotionally salient aspects of the *F*0 contour (Busso et al., 2009). Gross statistics such as the mean, maximum and minimum values, and the range of the *F*0 are found to be the most salient aspects of *F*0 contour. They also conduct their experiment by extracting features on the sen- tence and voiced regions levels. The results showed that features from the sentence level surpass the features from the voiced region level.

PERFORMANCES

The performance of the prosodic features based on their granularity is also analyzed in several studies. For instance, (Schuller et al, 2003)(Rao et al, 2013) showed that the performance is usually increased when global features are combined with local ones.

~~Schuller et al. compare gross statis- tics of pitch and energy contours, to instantaneous pitch and energy features using continuous Hidden Markov Model (Schuller et al., 2003). They obtained 86.6% recognition rate using global features, 77.6% by local ones while human judges have a recognition rate 79.8%. Rao et. al compared the local and global prosodic features, and their combination (Rao et al., 2013). The global features are computed from gross statistics of prosodic features. The local prosodic features are gathered from the sequence of syllable duration, frame level pitch and energy values. Compared to the performance of the local features, when the local and global prosodic features are combined, performance is slightly increased. It is also observed that from the word and syllable level prosodic analysis, final words of sentences and syllables involve more information to distinguish emotions compared to other parts of words and syllables.~~

**Spectral Features** TODO: per ora non ho messo una definizione di cosa sono, non è banale, che faccio ?

Spectral features are the characteristics of various sound components generated from different cavities of the vocal tract system. When sound is produced by a person, it is filtered by the shape of the vocal tract. The sound that comes out is determined by this shape. An accurately simulated shape may result in an accurate representation of the vocal tract and the sound produced. Characteristics of this cavity, the vocal tract, are well represented in the frequency domain (Koolagudi and Rao, 2012). Spectral features are obtained by transforming the time domain signal into the frequency domain using the Fourier transform. They are extracted from speech segments of length 20 to 30 milliseconds that are partitioned by a windowing method. Spectral features can be extracted in a number of ways including the ordinary **linear predictor coefficients (LPC)** [31], **one-sided autocorrelation linear predictor coefficients (OSALPC)** [32], **short-time coherence method (SMC)** [33], and **least-squares modified Yule–Walker equations (LSMYWE)** [34].   
However, to better exploit the spectral distribution over the audible frequency range, the estimated spectrum is often passed through a series of band-pass filters. Spectral features are then extracted from the outputs of these filters. **Mel Frequency Cepstral Coefficients (MFCC)** feature represents the short term power spectrum of the speech signal and it is the most widely used spectral feature (Kuchibhotla et al., 2014). **Linear Prediction Cepstral Coefficients(LPCC)** also embodies vocal tract characteristics of speakers. Those characteristics show differences with particular emotions. Another feature, **Log-Frequency Power Coefficients (LFPC)**, mimics logarithmic filtering characteristics of the human auditory system. **Gammatone Frequency Cepstral Coefficients (GFCC)** is also a spectral feature obtained by a similar technique of MFCC extraction. Finally **formants** are the frequencies of the acoustic resonance of the vocal tract. They are computed as amplitude peaks in the frequency spectrum of the sound. They determine the phonetic quality of a vowel, hence used for vowel recognition.

PERFORMANCES

~~Sato et al. use segmental MFCC features for speech emotion recog- nition (Sato and Obuchi, 2007). They labeled each frame using multi- template MFCC clustering. They compared the performance with prosody based algorithms using k-nearest neighbors and compared with conventional MFCC based algorithms using HMM. They achieved better performance using the new method.~~

~~Bitouk et al. introduced a new set of spectral features which are statistics of MFCC calculated over three phoneme type classes of interest-stressed and unstressed vowels, and consonants in the utterance (Bitouk et al., 2010). Compared to prosodic features or utterance level spectral features, they yielded results that have higher accuracy using the proposed features. In addition, combination of these features with prosodic features also increase accuracy. It has been also found that com- pared to stressed and unstressed vowel features, the consonant regions of the utterance involve more emotional information.~~

**Voice Quality Features**

Voice quality is determined by the physical properties of the vocal tract. Involuntary changes may produce a speech signal that might differentiate emotions using properties such as the **jitter**, **shimmer**, and **harmonics to noise ratio (HNR).   
Jitter** is the variability of fundamental frequency between successive vibratory cycles, while **shimmer** is the variability of the sound wave’s amplitude. Jitter is a measure of frequency instability, whereas shimmer is the amplitude instability. **Harmonics to Noise Ratio** is the measurement of the relative level of noise in the frequency spectrum of vowels. It is the ratio between periodic to aperiodic component in voiced speech signals. These variations are perceived as changes in voice quality.

Anyway, in the literature other voice quality measures can be found and generally speaking it can be said that the voice quality features are more auxiliary than primary features for a speech emotion recognition system (*M.B. Akçay and K. Oğuz*). For instance, as reported in (*A Survey of Classification Techniques in Speech Emotion Recognition)*, voice quality measures for a speech signal include harshness, breathiness, and tenseness. However, the relation of voice quality features with different emotions is not a deeply investigated area, and researchers have produced contradictory conclusions about this.

PERFORMANCES

*Lots of research (Lugger and Yang, 2007) (Li et al., 2007). (Zhang, 2008) showed that combining these features with prosody and/or spectral features increase the performances.*

***Teager energy operator (TEO) based features***

*There are features that depend on the Teager Energy Operator (TEO). It is used to detect stress in speech and has been introduced by Teager and Teager (1990) and Kaiser (1990, 1993). According to Teager, speech is formed by a non-linear vortex-airflow interaction in the human vocal system. A stressful situation affects the physiological muscle tension of the speaker that results in an alteration of the airflow during the sound’s emission. The operator developed by Teager to measure the energy from a speech by this non-linear process was documented by Kaiser as follows where Ψ[] is Teager Energy Operator and x(n) is the sampled speech signal.*

*Ψ[𝑋(𝑛)] = 𝑥^2(𝑛) − 𝑥(𝑛 + 1)𝑥(𝑛 − 1)*

*A feature that obtained a discrete success is the critical band based TEO auto-correlation envelope area (TEO-CB-Auto-Env). TEO-CB-Auto-Env outperforms both pitch and MFCC under stress condition (Zhou et al., 2001)*

*TODO: decidere se mettere una tabella riassuntiva, ma verrebbe enorme per cui non saprei. Io lascerei così e metterei la tabella solo per gli algoritmi perchè in ogni caso è veramente troppo variabile il tema, non si riesce a ottenere una tabella che abbia un ‘senso’*

*TODO: per ora non ho approfondito molto gli algoritmi: lo si farà una volta scelto su cosa sviluppare la tesi*

*2.4.2 Classifiers*

*The next step is to find a predictive function called predictor that associates each utterance to the conveyed emotion: this is done by learning. Different classification models take different approaches to learning. For SER, the prediction task is usually considered as a multiclass classification problem since each sample may belong to one of the emotions used as classes (if we refer to a categorical model). So, speech emotion recognition algorithms classify underlying emotions for a given utterance. Including traditional classifiers and deep learning algorithms, many machine learning algorithms are used to carry out this speech emotion recognition task.  
Here I reported**a list of classifiers commonly used in SER along with literature references.*

|  |  |
| --- | --- |
| ***Classifier*** | ***References (TODO)*** |
| ***Hidden Markov model*** | ***HMM*** |
| ***Gaussian mixture model*** | ***GMM*** |
| ***K -Nearest neighbor*** | ***KNN*** |
| ***Support vector machine*** | ***SVM*** |
| ***Artificial neural network*** | ***ANN*** |
| ***Bayes classifier*** | ***Bayes*** |
| ***Linear discriminant analysis*** | ***LDA*** |
| ***Deep neural network*** | ***RNN + CNN*** |
| ***Decision Tree*** | ***Decision Tree*** |

*Although in the table there are nine classifiers listed, the most ‘successful’ and prominent ones are HMM, GMM, SVM and DNN.  
 In the following I reported a brief intuition of what is the idea behind these techniques, but first It is worth mentioning that it is really hard to properly describe these algorithms in terms of performances in the SER field. This is because the researches and studies carried out are often unrelated: they often refer to different datasets (this may introduce different kind of biases), they use different combinations of features, different models of emotions or different number of categories (classes) when adopting categorical models and so on. Also, sometimes they use speaker dependent data and sometimes they don’t: when a speaker is known to a listener, the recognition is treated as speaker dependent, that is, the prior knowledge about the speaker is taken into account during the recognition process. This speaker dependence obviously improves humans’ recognition performance. So, unfortunately, as already pointed out, the performances of the various algorithms are not properly comparable. Anyway, the performances reported refer to the average Acccuracy of the algorithms: Accuracy = Number of correct predictions / Total number of predictions.*

*Hidden Markov Model is a commonly used method for speech recognition and has been successfully extended to recognize emotions, as well. As the name suggests, HMM relies on the Markov property which says that the current state of a system at a time t only depends on the previous state at time 𝑡 − 1 so that we can forget the ‘history’. The term “hidden” denotes the inability of seeing the process that generates the state at time t. It is then possible to use probability to predict the next state by making observations of the current state of the system. HMMs are suitable for the sequence classification problems that consist of a process that unfolds in time. That is why HMM is very successful in speech recognition systems where the sequence of the spoken utterance is a time-dependent process. The HMM parameters are tuned in the model training phase to best relate the training data to the known category. The model then classifies unseen data based on the highest posterior probability.*

*Performances:*

*HMM 1: emotions = big six + neutral, They used energy and pitch features and with their best possible combination and the number of HMM states they obtained an accuracy for all the emotions higher than 70% and an average accuracy of 82.5%*

*HMM 2: emotions = big six + neutral. They used acted emotions (by non professionals I think). They obtained an average overall accuracy of 77.8% by using continous HMM using low-level instantaneous features (to exploit the warping capabilities of HMMs “global statistics are clearly invalid”). In the same research they also used GMM with global statistics obtaining better performances, however, for what concerns HMM, this research suggests that low-level features prove their full potential only when using higher number of states.*

*In both researches however a general increase in performance can be observed using more states.*

*HMM 3: emotions = anger, happiness, sadness, surprise + neutral, they used the DES database. Features based on the fundamental frequency, energy, formants, and Mel sub-band energies were extracted as the candidate input to the HMM classifier. The best feature set was selected by the SFS method and an average accuracy of 99.5% was obtained for the gender independent case.   
In the SVM based system, the Mel energy spectrum dynamics coefficients were utilized to classify the five emotional states and obtained an average accuracy of 88.9%*

*TODO: aggiungi INTERFACE ai databases*

*SER researchers used HMM for a long time and used it with various types of feature sets. For example, some researchers used prosody features, and some used spectral features. Researchers using the HMM achieved promising average SER classification accuracies comparable with other classification techniques, but further improvement possibilities are low [A Survey of Classification Techniques in Speech Emotion Recognition]: that is why HMM has been replaced by other classification techniques in later studies such as SVM, GMM, or DNN.*

*Gaussian Mixture Model Gaussian Mixture Model is a probabilistic method which can be viewed as a special case of continuous HMM that contains only one state. It is based on the idea that an unknown distribution p(x) can be described by a convex combination of N base distributions like Gaussians: GMM is the special case of mixture models, where the base distribution is assumed to be Gaussian. So, the idea behind the mixture models is modeling the data in terms of a mixture of several components, where each component has a simple parametric form, a Gaussian in this case. It is assumed that each data point belongs to one of the components, and it is tried to infer the distribution for each component separately.*

*Performances*

*GMM 1: emotions = big six + neutral and an average accuracy of 86.8% by using Gaussian mixture models using derived global features of the raw pitch and energy contour of the speech signal. They used acted emotions (by non professionals I think).*

*GMM 2: emotions = anger, boredom, disgust, fear, joy, sadness + neutral (Berlin database). This works points out that a lot of works didn’t take into account the fact that databases may be unbalanced (for instance I may have more utterances for one emotion than another): this characteristic of the database can bias the validation of methods and produce biased and non-comparable results. In this work they balanced the database (same number of samples for each emotion) and obtained an average recognition rate (acccuracy) of 63.49% for GMM of 68.57% for HMM, but higher accuracy (71.75%) was obtained by using hierarchical models anyway. The features used were: MLS, MFCCs and prosodic features.*

*So, GMM is one of the most popular classification algorithms among SER researchers, and many research works are based on GMM. GMM classifiers have least training time among the prominent classifiers and this made it an attractive choice as SER classifier [A Survey of Classification Techniques in Speech Emotion Recognition]. When global features are extracted from speech for SER, fewer data points are available, but GMMs work better in those scenarios and the average training time is minimal for GMM [8].*

*Support Vector Machine SVMs are supervised classifiers which find an optimal hyperplane for linearly separable data points. Given the training data the objective of an SVM classifier is to find the hyperplane that has the maximum margin from the closest data points of both classes. If the original data points are not linearly separable, it is possible to use a kernel function to map the data into a new space.*

*Performances:*

*SVM2: emotions = disgust, boredom, neutral, happiness and sadness, they used Berlin emotional database. The features extracted from these utterances are energy, pitch, linear prediction cepstrum coefficients (LPCC), Mel Frequency cepstrum coefficients (MFCC), Linear Prediction coefficients and Mel cepstrum coefficients (LPCMCC). The Support Vector Machine (SVM) is used as a classifier to classify different emotional states. The system gives 66.02% classification accuracy for only using energy and pitch features, 70.7% for only using LPCMCC features, and 82.5% for using both of them.*

*SVM3: Berlin EMO DB, CASIA DB, Chinese elderly emotion database(EESDB). They used Fourier Parameters, MFCC. For EMO DB they used Happ. Bored. Neutr. Sad. Angry Anxi. and obtained 88.88% recognition rate, while for CASIA DB with Happ.Surpri Neutr. Sadn. Anger Fear, they reached 79% recognition rate, and finally for EESDB by using only Happ. Sadn. Anger Neutr., they obtained a 76% recognition rate.  
  
~~SVM is extensively used in SER [48, 49, 55, 65–71]. Performance of SVM for SER task in most of the research studies carried out yielded nearly close results, and accuracy is varying around 80% mark. However, Hassan and Damper [69] achieved 92.3% and 94.6% classification accuracy using linear and hierarchical kernels, respectively. They have used a linear kernel instead of a nonlinear radial basis function (RBF) kernel because of very high dimensional features space [72]. Hu et al. [73] explored GMM supervector-based SVM with different kernels like linear, RBF, polynomial, and GMM KL divergence and found that GMM KL performed the best in classifying emotions.~~*

*Most of the deep learning algorithms are based on Artificial Neural Networks hence are commonly referred to deep neural networks. Artificial Neural Networks are a commonly used method for a lot of classification problems. They are constructed with an input layer, one or more hidden layers, and an output layer. Each layer is made up of nodes: the number of nodes in the input and output layers depends on the representation of the data and the labeled classes, while the hidden layers can have as many nodes as required. Each layer is connected to the next using weights that are initially randomly chosen. The samples’ values are loaded into the input layer, and then forwarded to the next layer. At the output layer, the weights are updated at each iteration using the backpropagation algorithm. Once the training is complete, the weights are expected to be able to classify new data.*

*Performances*

*NN 1: emotions = big six + neutral, they acquired their own acted data in a lab (not with professional actors I think).. They used a 33 dimensional acoustic feature set and they obtained an accuracy of 74.83% for GMM, of 76.12% for SVM and of 73.15% for a neural network that however have good performances on small training sets compared to GMMs as an advantage.*

*The term “deep” comes from the number of inner layers that may reach hundreds (while in a traditional neural network the number of hidden layers is of two or three). Convolutional neural networks (CNNs) [110, 111] and recurrent neural networks (RNNs) [112] are the most successful DL algorithms for SER. Convolutional networks are neural networks that use convolution in place of general matrix multiplication in at least one of their layers, whereas when feedforward neural networks are extended to include feedback connections, they are called RNNs. RNNs are specialized in processing sequential data.*

*Performances*

*CNN 1: SAVEE DB (, anger, disgust,fear, happiness, sadness, surprise, and neutral), Berlin EMO DB (, anger, disgust, fear, joy, sadness, boredom and neutral), DES DB (anger, joy, surprise, sadness and neutral), MES DB (anger,joy, surprise, sadness and disgust) and they obtained respectively an accuracy of 73.6%, 85.2%, 79.9% and 78.3%.*

*RNN 2: here they studied the use of deep learning to automatically discover emotionally relevant features from speech (both the low level descriptors and high level statistical functions). They used the IEMOCAP dataset that is an acted one (professional actors I think) and categorize the emotions in only four classes: angry, happy, neutral, sad. The designed system has an average accuracy of 61.8% when using raw spectral features and a 63.5% accuracy with low level descriptors typically used in SER (fundamental frequency (F0), voicing probability, frame energy, zero-crossing rate, and 12 Mel-frequency Cepstral Coefficients (MFCC)). They also used an SVM classifier as baseline on the same dataset: the learned features provided better classification accuracy compared to this traditional SVM-based SER using fixed designed features.*

*~~Researchers are applying end-to-end DL systems in SER [84, 91–94], and most of them use dimensional models of emotions. Although using end-to-end DL, the average classification accuracy for arousal is 78.16%, for valence it is pretty low: 43.06%. Among other DNN techniques, very recently, a maximum accu- racy of 87.32% is achieved by using a fine-tuned Alex-Net on Emo-DB [85]. Han et al. [113] used extreme learning machine (ELM) for classification, where a DNN takes as input the popular acoustic features within a speech seg- ment and produces segment-level emotion state probability distributions, from which utterance-level features are constructed.~~ In the last years, performances of the deep learning algorithms surpassed the traditional machine learning algorithms (M.B. Akçay and K. Oğuz ). The advantage of some of these algorithms is that there is no need for feature extraction and feature selection steps: all the features are automatically selected with deep learning algorithms*

*TODO: rimuovi questa tabella o al limite fanne una in cui riassumi gli studi che hai descritto*

*~~In the following table a coarse recap of the performances of the most prominent algorithms is reported.~~*

|  |  |
| --- | --- |
| ***~~Classifier~~*** | ***~~Average accuracy~~*** |
| ***~~Hidden Markov model~~*** | ***~~75.5%-78.5%~~*** |
| ***~~Gaussian mixture model~~*** | ***~~74.83%-81.94%~~*** |
| ***~~Support vector machine~~*** | ***~~around 80%~~*** |
| ***~~Deep neural network~~*** | ***~~78.16% for Arousal~~***  ***~~43.06% for Valence~~*** |

*TODO: DELETE AFTERWARDS (it contains the definitions of the algorithms, it may be useful)*

***Traditional Classifiers***

*SER systems typically make use of classification algorithms. A classification algorithm requires an input X, an output Y, and a function that maps X to Y as in 𝑓(𝑋) = 𝑌 . The learning algorithm approximates the mapping function, which helps predict the class of new input. The learn- ing algorithm needs labeled data which identifies the samples and their classes. Once the training is over, data that has not been used during training is used to test the performance of the classifier. Most preferred algorithms are Hidden Markov Model (HMM), Gaussian Mixture Model (GMM), Support Vector Machines (SVM), and Ar tificial Neural Networks (ANN). There are also classification methods based on Decision Trees (DT), k-Nearest Neighbor (k-NN), k-means, and Naive Bayes Classifiers. In addition to usage of single classifiers, ensemble methods are also used for SER that combines several classifiers to obtain better results.*

*Hidden Markov Model. Hidden Markov Model is a commonly used method for speech recognition and has been successfully extended to recognize emotions, as well. As the name suggests, HMM relies on the Markov property which says that the current state of a system at a time t only depends on the previous state at time 𝑡 − 1. The term “hidden” denotes the inability of seeing the process that generates the state at time t. It is then possible to use probability to predict the next state by making observations of the current state of the system.*

*Gaussian Mixture Model Gaussian Mixture Model is a probabilistic method which can be viewed as a special case of continuous HMM that contains only one state. The idea behind the mixture models is modeling the data in terms of a mixture of several components, where each component has a simple parametric form, such as a Gaussian. It is assumed that each data point belongs to one of the components, and it is tried to infer the distribution for each component separately.*

*Artificial Neural Networks. Artificial Neural Networks are a commonly used method for several kinds of classification problems. It is basically constructed with an input layer, one or more hidden layers, and an output layer. The layers are made up of nodes; while the number of nodes in the input and output layers depends on the representation of the data and the labeled classes, the hidden layers can have as many nodes as required. Each layer is connected to the next using weights that are initially randomly chosen. When a sample is chosen from the train- ing data, its values are loaded to the input layer, and then forwarded to the next layer. At the output layer, the weights are updated using the backpropagation algorithm. Once the training is complete, the weights are expected to be able to classify new data.*

*Support Vector Machine SVMs are supervised classifiers which find an optimal hyperplane for linearly separable patterns. Given the training data the objective of an SVM classifier is to find the hyperplane that has the maximum margin, between data points of both classes. If these patterns are not linearly separable, using a kernel function original data points are mapped to a new space.*

*Ensemble of Classifiers. In ensemble learning, a number of ma- chine learning algorithms are combined to increase predictive perfor- mance. Each algorithm in ensemble classifier is combined in some way, typically by a voting procedure, to obtain a final result. Performances of the ensembles are often higher than the individual classifiers. Different types of architectures are available in ensemble classifiers. One of the ways is feeding the same data to each classifier by comparing the results obtaining a final decision. Another approach is using the hierarchical classifier. In this approach, input data is fed to one algorithm, then the result is fed to another type of classifier in a hierarchical approach, then the final decision is given.*

***Deep learning based classifiers***

*RNNs are a family of neural net- works which are specialized in processing sequential data. By the usage of internal memory, they can remember the received input data and make a precise prediction about what is coming next. Because of their nature, RNNs are successfully used for sequential data such as time se- ries, speech, text, video. When a unit of RNN produces an output, it forwards data to the next unit and also loops the output back itself. As a result, it has two types of input: present input and input from the recent past. The input from the recent past is important because the sequence of the data contains important information about what is coming next.*

*RNNs have a short time memory, however, by using Long-Short Time Memory architecture, RNN can gain access to long term memory. LSTM- RNNs are a kind of gated RNN which are the most effective models used in practical applications that solves the long term dependency problem of the RNN. LSTM-RNNs have special “LSTM cells” that have internal recurrence besides the outer recurrence of RNN. In addition to standard input and output of the RNN, it has more parameters and gating units with sigmoid nonlinearity that control the flow of information. LSTM has three types of gates: input gate, forget gate and remember gate. By opening and closing these gates, LSTM cell makes decisions about what to store, and when to allow inputs, outputs, and deletions.*

*Convolutional Neural Networks. Convolutional Neural Net-works (CNNs) are particular types of neural networks which are de- signed to process data that has a grid-like topology, such as images. Through applications of several relevant filters, CNN can successfully capture temporal and spatial dependencies from an input source. The inputs are reduced into a form without loss of feature so that compu- tational complexity decreases and the success rate of algorithm is in- creased. A CNN is composed of several layers: convolution layer, polling layer, and Fully-Connected layer. A convolution layer is used to extract high-level features from the input. Mathematically a convolution means combining two functions to obtain a third one. In CNN, the input is taken and, then a kernel is applied to it. The resulting output is a feature map. Polling layer is used to reduce the size of convoluted features to de- crease computational complexity through dimensionality reduction. It is useful for extracting dominant features of the input data. After passing input from several convolution and polling layers and extracting the high-level features, the resulting features are used as an input to a fully connected layer by flattening the 2D data to a column array and feeding it to a feed-forward network that operates as an ordi- nary neural network.*

1. Thesis Development

Packages installed: opensmile, scikitlearn

3.1 Feature Extraction

The data features used are the ones of GeMAPS [cit]: TODO parla di queste features guardando il paper → riportale tutte e descrivele e motiva il perchè le hai usate (sono le cose sottolineate)  
In order to extract the chosen features the OPENSMILE[cit] tool was used: TODO parlane  
TODO parla del preprocessing dei dati dopo aver estratto le features

3.2 Algorithms

General Preprocessing:

Performed normality test on the EMOVO dataset and verified that they’re not distributed as a gaussian, so it is better to normalize (MinMaxScaler) than standardize. <https://stackoverflow.com/questions/30918781/right-function-for-normalizing-input-of-sklearn-svm>

*SVM*[*https://www.datacamp.com/community/tutorials/svm-classification-scikit-learn-python*](https://www.datacamp.com/community/tutorials/svm-classification-scikit-learn-python)

*SVM is an exciting algorithm and the concepts are relatively simple. The classifier separates data points using a hyperplane with the largest amount of margin. That's why an SVM classifier is also known as a discriminative classifier. SVM finds an optimal hyperplane which helps in classifying new data points. SVM generates optimal hyperplane in an iterative manner, which is used to minimize an error. The core idea of SVM is to find a maximum marginal hyperplane(MMH) that best divides the dataset into classes.*

*Support vectors are the data points, which are closest to the hyperplane. These points will define the separating line better by calculating margins. These points are more relevant to the construction of the classifier.*

*A hyperplane is a decision plane which separates between a set of objects having different class memberships.*

*A margin is a gap between the two lines on the closest class points. This is calculated as the perpendicular distance from the line to support vectors or closest points. If the margin is larger in between the classes, then it is considered a good margin, a smaller margin is a bad margin.*

## ***How does SVM work?***

*The main objective is to segregate the given dataset in the best possible way. The distance between the either nearest points is known as the margin. The objective is to select a hyperplane with the maximum possible margin between support vectors in the given dataset. SVM searches for the maximum marginal hyperplane in the following steps:*

1. *Generate hyperplanes which segregates the classes in the best way. Left-hand side figure showing three hyperplanes black, blue and orange. Here, the blue and orange have higher classification error, but the black is separating the two classes correctly.*
2. *Select the right hyperplane with the maximum segregation from the either nearest data points as shown in the right-hand side figure.*

#### ***Dealing with non-linear and inseparable planes***

*Some problems can’t be solved using linear hyperplane, as shown in the figure below (left-hand side).*

*In such situation, SVM uses a kernel trick to transform the input space to a higher dimensional space as shown on the right. The data points are plotted on the x-axis and z-axis (Z is the squared sum of both x and y: z=x^2=y^2). Now you can easily segregate these points using linear separation.*

*Steps followed:*

*Paper -> SVM 3*

*preprocessing-> Normalization is an important aspect for a robust emotion recogni- tion system [30]. The goal is to eliminate speaker and recording variability while keeping the effectiveness of emotional discrimina- tion [51]. In particular, it could compensate for speaker variability. Here, z-score normalization [51] was adopted for feature normalization. For a given FP feature H from a speech signal of a speaker s, its mean value, EðHsÞ, and its standard deviation, std(Hs), were first derived out. Then, the normalized feature was estimated by the fol- lowing (3).*

***scaling each attribute*** *in the dataset. If we don’t do it, classifier can depend much more on attributes which scale is larger than others. Here, for example, R, G, B color values are greater than relative segment position values. To prevent it, one could perform linear scaling which can result in e.g. 0 to 1 range. It’s based on maxima and minima of a given attribute in it’s whole population. But it’s not good idea when we expect to have any outliers. Let’s try different approach. We can center each value against the calculated mean. But still, ranges are different for various features. So, we divide the result by the standard deviation of the population. This method is called a Z-score (standard score) scaling. It’s calculated from the following formula:*

*z = (X – µ) / σ*

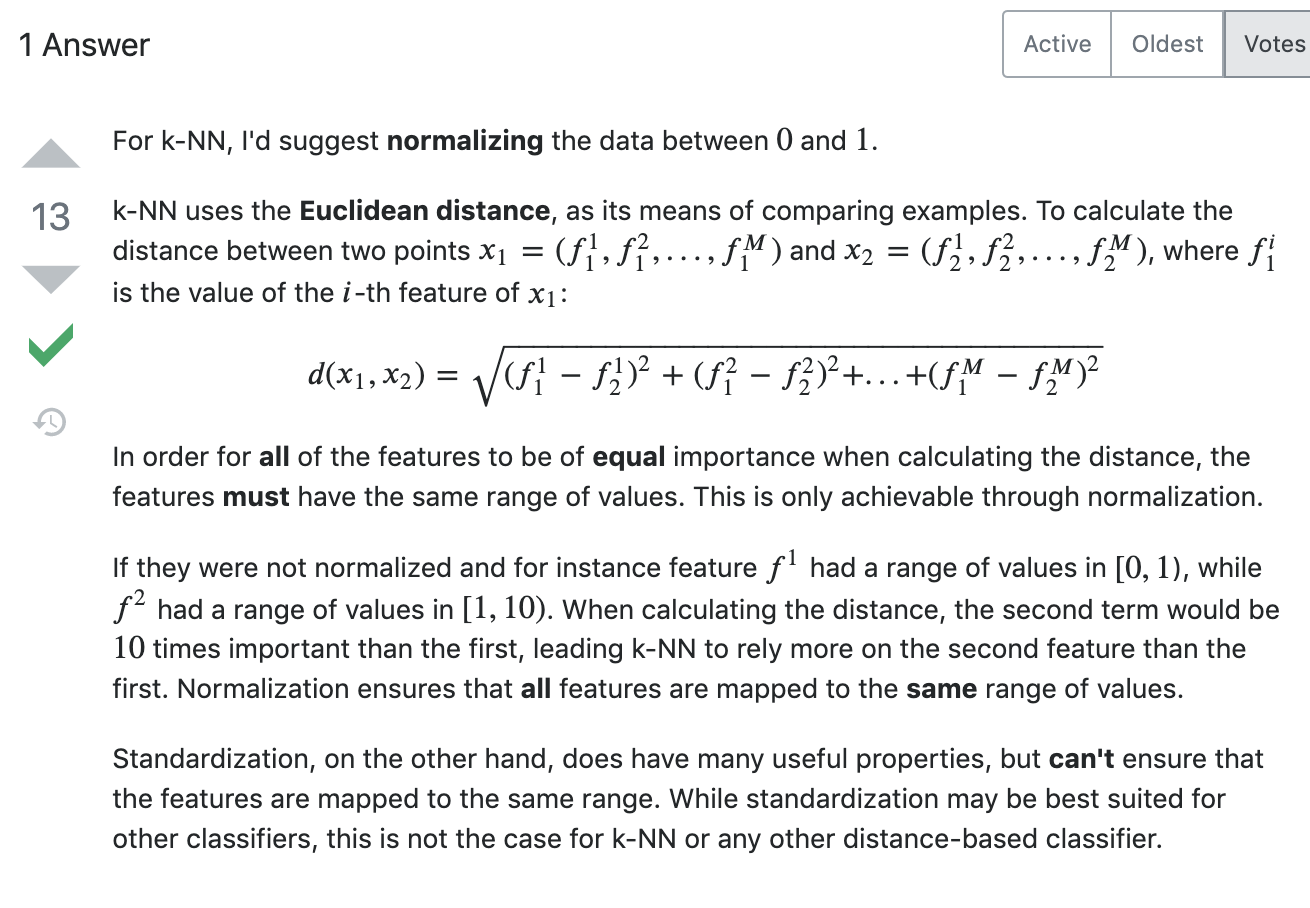
*where z is the z-score, X is the value of the element, µ is the population mean, and σ is the standard deviation.*

*Discorso su data leakage: posso fare il preprocessing prima o dopo lo split in training and test set.*

*Discorso su stratify per avere sets (training and test) bilanciati*

*KNN*

*Preprocessing:*

**

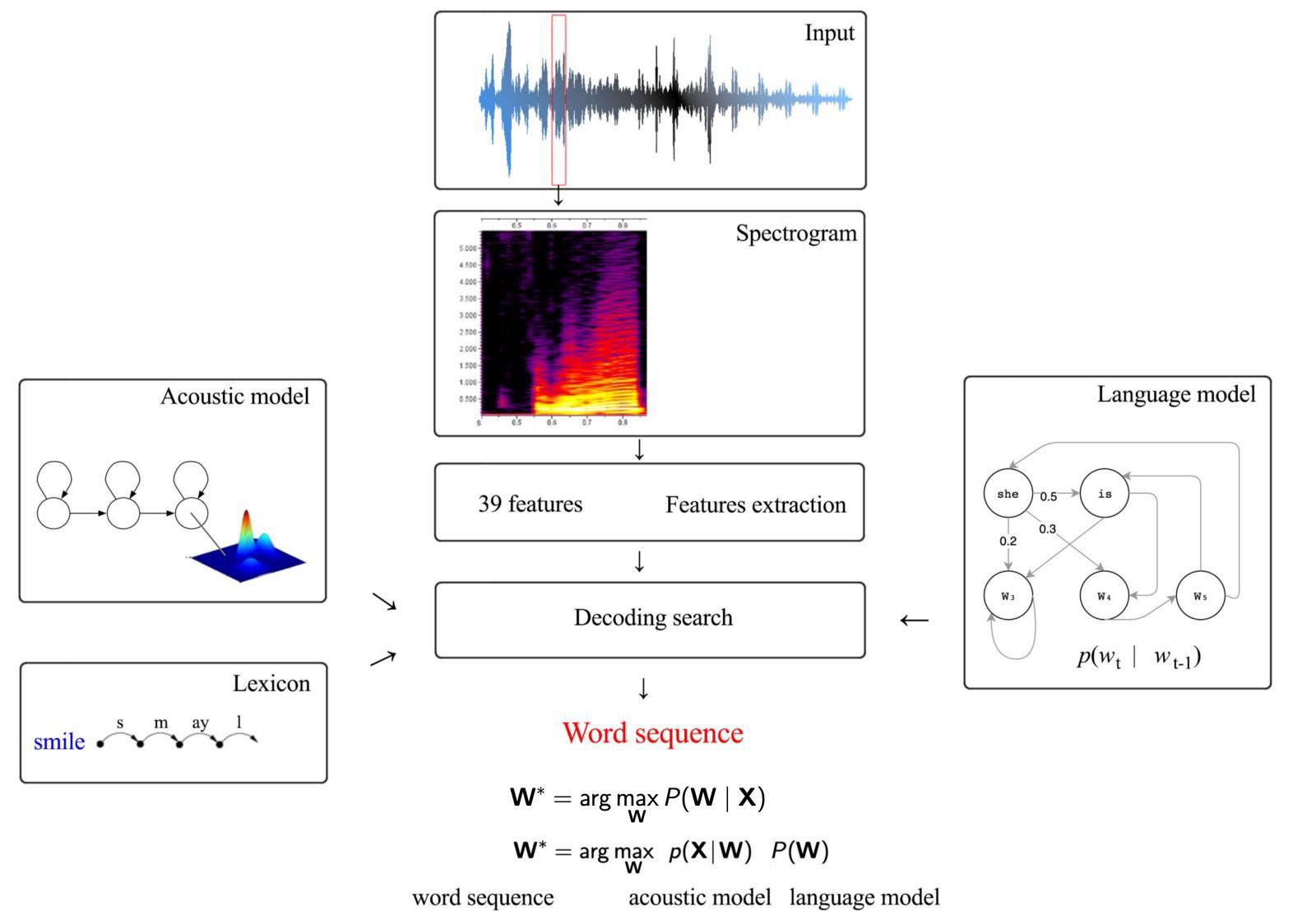
*HMM*

*HMM can be used for classification. This is a straightforward application of the bayesian classification framework, with the HMM being used as the probabilistic model describing your data. For example, you have a large database of utterances of digits ("one", "two", etc) and want to build a system capable of classifying an unknown utterance. For each class in your training data ("one", "two", you estimate the parameters of a HMM model describing the training sequences in this class - and you end up with 10 models. Then, to perform recognition, you compute the 10 likelihood scores (which indicate how likely the sequence you want to recognize has been generated by the model), and the model with the highest score gives you the digit. In the* [*Rabiner tutorial on HMMs*](http://www.cs.cornell.edu/Courses/cs4758/2012sp/materials/hmm_paper_rabiner.pdf)*, the training stage is "Problem 3", the classification stage is "Problem 2".*

# *Automatic Speech Recognition (ASR)*

*Let’s get a high-level overview first. The diagram below is a high-level architecture for speech recognition that links HMM (Hidden Markov Model) with speech recognition.*

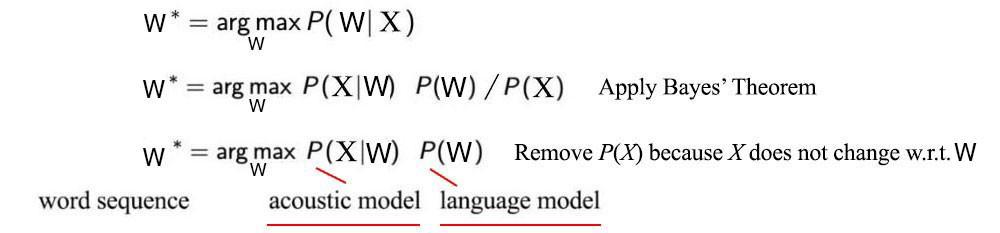
**

**

*Starting from an audio clip, we slide windows of 25 ms width and 10 ms apart to extract* [*MFCC features*](https://medium.com/@jonathan_hui/speech-recognition-feature-extraction-mfcc-plp-5455f5a69dd9)*. For each window frame, 39 MFCC parameters will be extracted. The primary objective of speech recognition is to build a statistical model to infer the text sequences W (say “cat sits on a mat”) from a sequence of feature vectors X.*

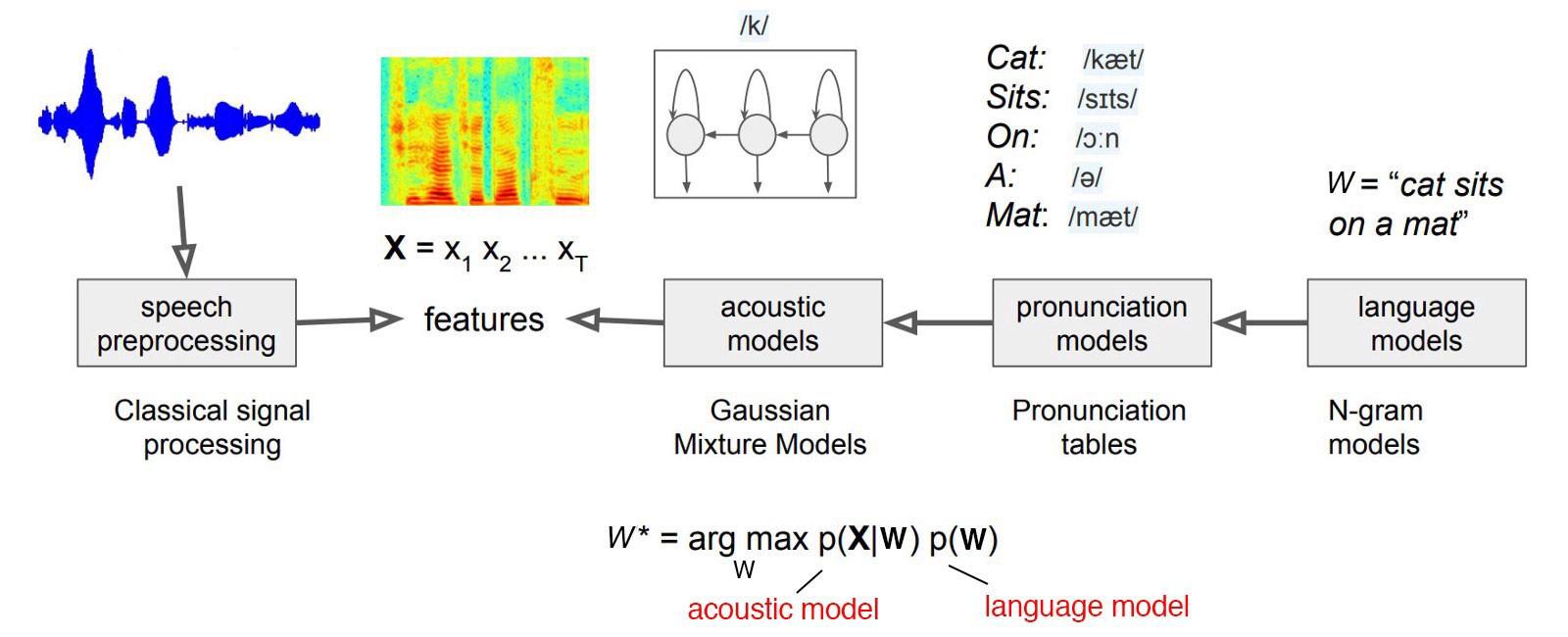
*One approach looks for all possible sequences of words (with limited maximum length) and finds one that matches the input acoustic features the best.*

**

**

*This model depends on building a language model P(W), a pronunciation lexicon model, and an acoustic model P(X|W) (a generative model) as below.*

**

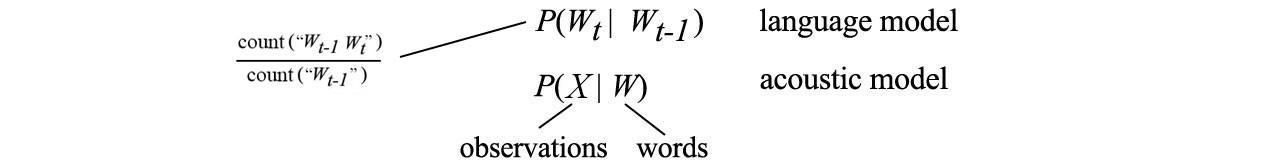
**

*Modified from* [*source*](https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1174/lectures/cs224n-2017-lecture12.pdf)

*A pronunciation model can use tables to convert words to phones, or a corpus is already transcribed with phonemes already. The acoustic model is about modeling a sequence of feature vectors given a sequence of phones instead of words. But we will continue the use of the notation p(X|W) for the acoustic model. Just be aware.*

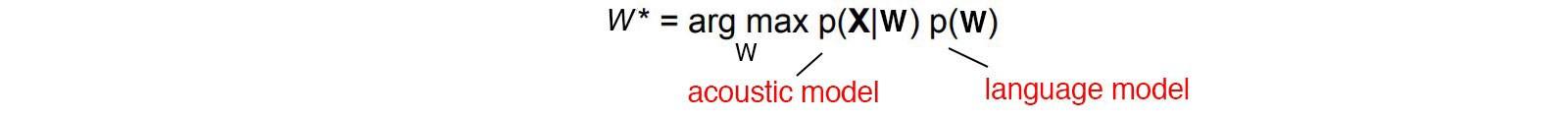
*The language model is about the likelihood of the word sequence. For example, “I watch a movie” will be more likely than “I you movie watch” or “I watch an apple”. It predicts the next word given the previous words. If we approximate it with a first-order Markov chain, the next word will depend on the current word only. We can estimate it by counting the occurrence of word pairs in a corpus.*

**

**

*By combining the acoustic model and the language model, we search for the text sequence with the maximum likelihood.*

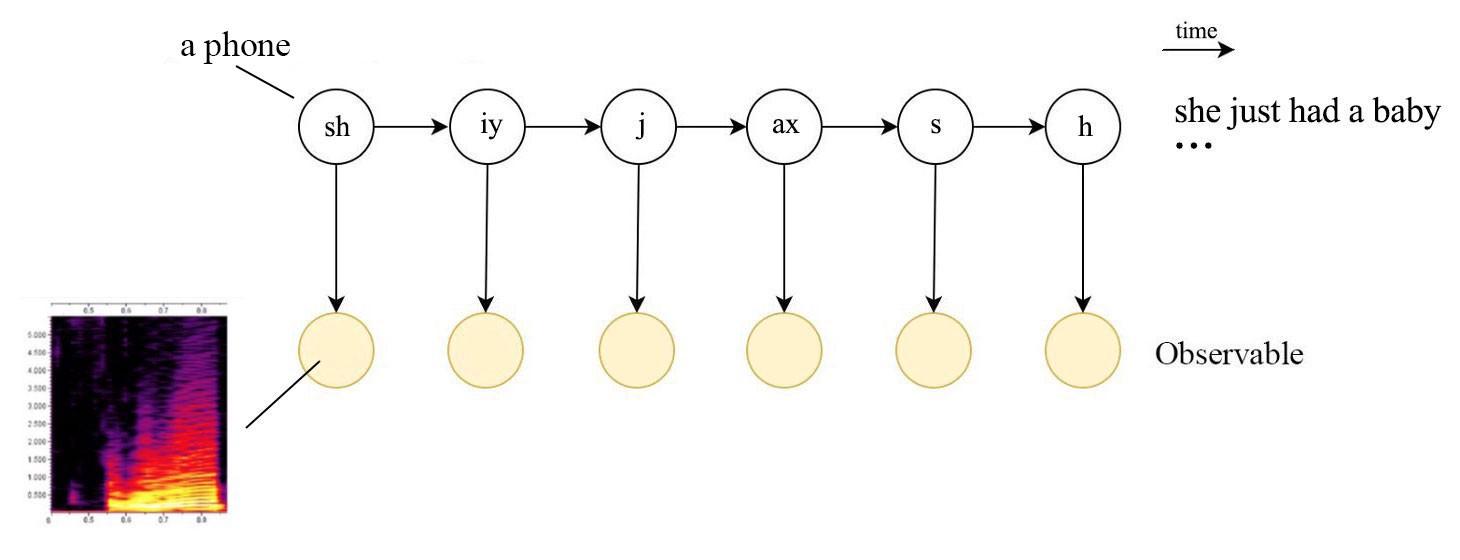
**

**

*This approach sounds indirect and the search looks inefficient or impossible. But p(X|W) is much easier to model in speech recognition. The distribution of features for a phone can be modeled with a Gaussian Mixture Model (GMM). We will learn it with training data. The transition between phones and the corresponding observable can be modeled with the Hidden Markov Model (HMM). So if we can find an optimal way to search the phone sequence efficiently, this may not sound bad after all.*

*An HMM model composes of hidden variables and observables. The top nodes below represent the phones and the bottom nodes represent the corresponding observables (the audio features). The horizontal arrows demonstrate the transition in the phone sequence for the true label “she just …”.*

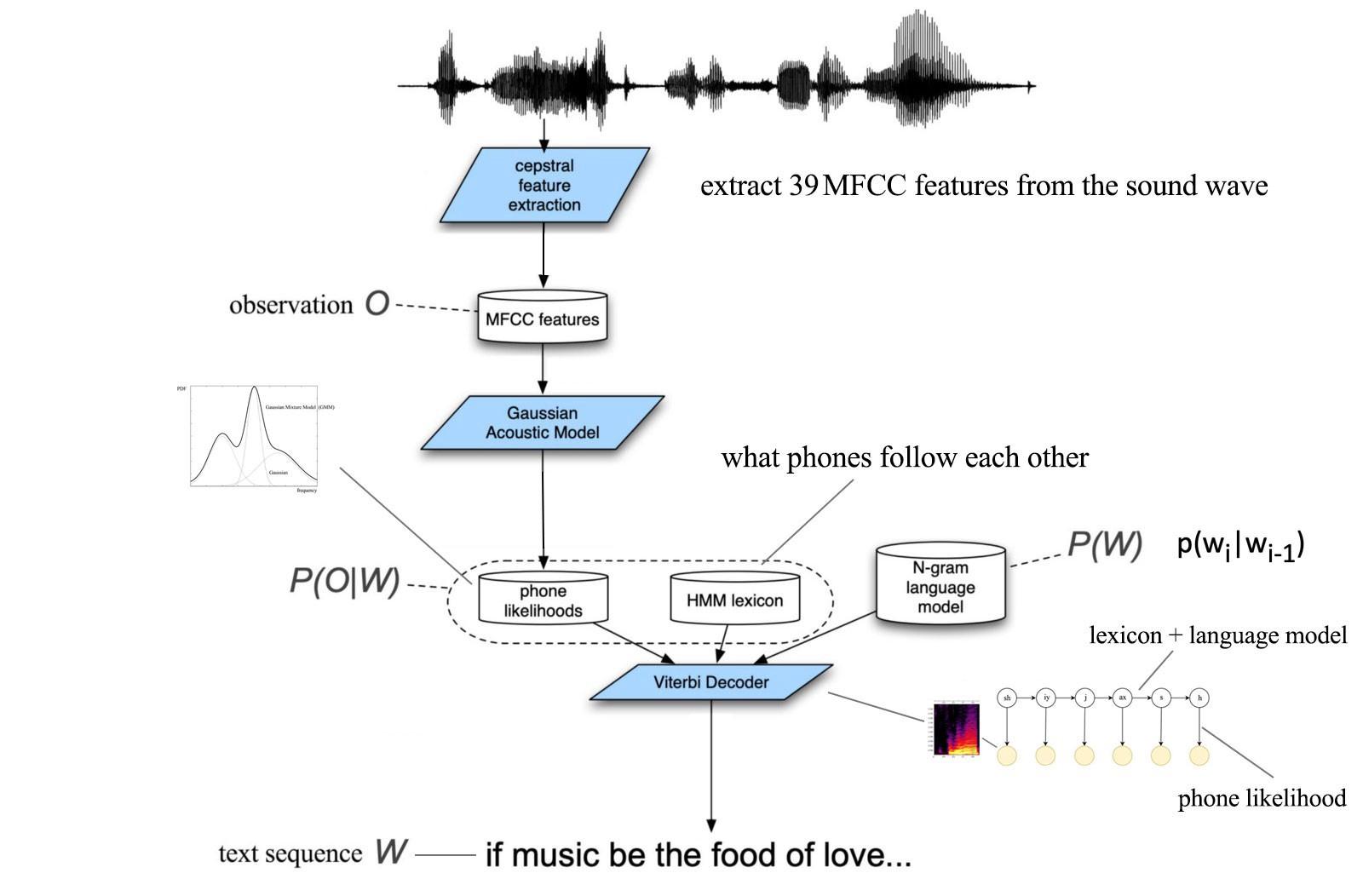
**

**

*In speech recognition, the observable can be represented by 39 MFCC features extracted from the corresponding audio frame. The good news is that with this HMM model, we do not need to search the phone sequence one-by-one. Otherwise, the complexity grows exponentially with the number of phones. With the Viterbi algorithm or other HMM methods, we can find the optimal sequence in polynomial time. We will come back to this later.*

*The diagram below is a possible realization of an Automatic Speech Recognition (ASR). Combining information on the lexicon, the acoustic model and the language model, we can find the optimal phone sequence with the Viterbi decoder.*

**

**

*Modified from* [*source*](https://web.stanford.edu/class/cs224s/lectures/224s.17.lec3.pdf) *(O is the same as X here)*

*Let’s do a quick recap, we can model the acoustic model P(X|W) with an HMM. The arrows on an HMM model will represent phone transitions or links to observables. To model the audio features that we observe, we learn a GMM model from the training data. So let’s understand HMM and GMM more in the general context first.*

*[From Schuller’s paper]*

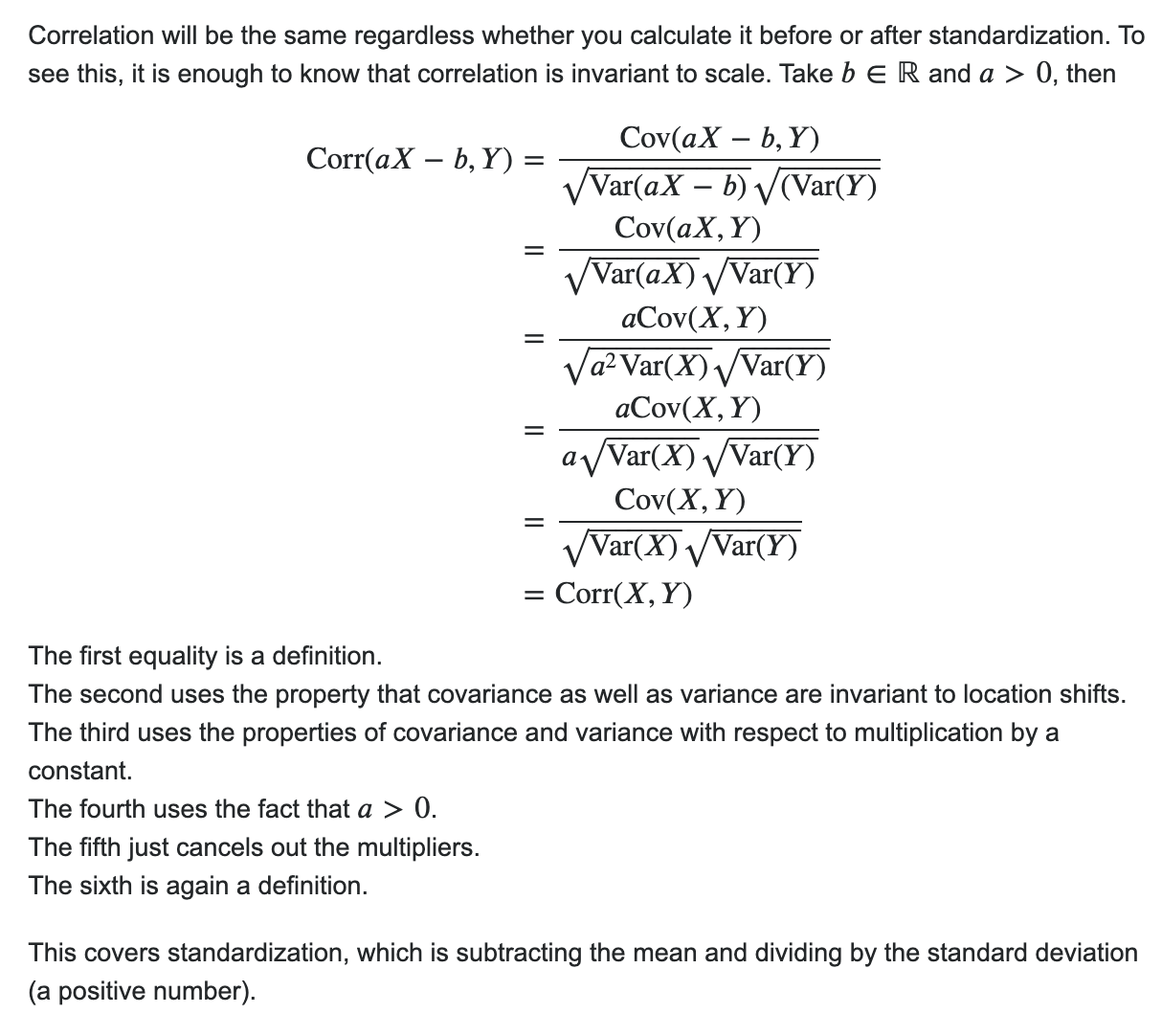
*A global statistics framework of an utterance is classified by Gaussian mixture models using the functional features of opensmile: this is ok in case of a one-state HMM (GMM), but won’t suits the case in which we have more than one state in the HMM (local features must be used in that case). They are classified by single state HMM’s (GMM), which are able to approximate the probability distribution function of each derived feature by means of a mixture of Gaussian distributions. Each emotion is modeled by one GMM in our approach. The maximum likelihood model will be considered as the recognized emotion at a time throughout the recognition process.*

*[From the paper Tuning Hidden Markov Model for Speech Emotion Recognition ]*

*Acoustic models are generated by one or few states HMMs. This states are associated with an emission- probability P (X |s) which for continuous variables x is replaced with its probability density function (PDF). These PDFs are realized using weighted sums of elementary Gaussian PDFs (Gaussian Mixtures, which leads to the name GMM). The GMM is a weighted sum of M component densities.*

*The Gaussian mixture model is parameterized by the mean vectors, covariance matrices and mixture weights from all component densities.*

*Correlation of the features*

*NB: *

**Data Correlation:** Is a way to understand the relationship between multiple variables and attributes in your dataset. Using Correlation, you can get some insights such as:

* One or multiple attributes depend on another attribute or a cause for another attribute.
* One or multiple attributes are associated with other attributes.

## So, why is correlation useful?

* Correlation can help in predicting one attribute from another (Great way to impute missing values).
* Correlation can (sometimes) indicate the presence of a causal relationship.
* Correlation is used as a basic quantity for many modelling techniques

Let’s get a closer look at what this means and how correlation can be useful. There are three types of correlations:

**Positive Correlation:** means that if feature **A** increases then feature **B** also increases or if feature **A** decreases then feature **B** also decreases. Both features move in tandem and they have a linear relationship.

**Negative Correlation:** means that if feature **A** increases then feature **B** decreases and vice versa.

**No Correlation:** No relationship between those two attributes.

Each of those correlation types can exist in a spectrum represented by values from 0 to 1 where slightly or highly positive correlation features can be something like 0.5 or 0.7. If there is a strong and perfect positive correlation, then the result is represented by a correlation score value of 0.9 or 1.

If there is a strong negative correlation, it will be represented by a value of -1.

## How Can I Deal With This Problem?

There are multiple ways to deal with this problem. The easiest way is to delete or eliminate one of the perfectly correlated features. Another way is to use a dimension reduction algorithm such as [Principle Component Analysis (PCA)](https://en.wikipedia.org/wiki/Principal_component_analysis).

*.*