Extracting Emotions and Communication Styles from Vocal Signals

Licia Sbattella, Luca Colombo, Carlo Rinaldi, Roberto Tedesco, Matteo Matteucci
and Alessandro Trivilini
Politecnico di Milano, Dip. di Elettronica, Informazione e Biongegneria, P.zza Leonardo da Vinci 32, Milano, Italy

Keywords: Natural Language Processing, Communication Style Recognition, Emotion Recognition.

Abstract: Many psychological and social studies highlighted the two distinct channels we use to exchange information among us—an explicit, linguistic channel, and an implicit, paralinguistic channel. The latter contains information about the emotional state of the speaker, providing clues about the implicit meaning of the message. In particular, the paralinguistic channel can improve applications requiring human-machine interactions (for example, Automatic Speech Recognition systems or Conversational Agents), as well as support the analysis of human-human interactions (think, for example, of clinic or forensic applications). In this work we present PrEmA, a tool able to recognize and classify both emotions and communication style of the speaker, relying on prosodic features. In particular, communication-style recognition is, to our knowledge, new, and could be used to infer interesting clues about the state of the interaction. We selected two sets of prosodic features, and trained two classifiers, based on the Linear Discriminant Analysis. The experiments we conducted, with Italian speakers, provided encouraging results (Ac=71% for classification of emotions, Ac=86% for classification of communication styles), showing that the models were able to discriminate among emotions and communication styles, associating phrases with the correct labels.

1 INTRODUCTION

Many psychological and sociological studies highlighted the two distinct channels we use to exchange information among us—a *linguistic* (i.e., explicit) channel used to transmit the contents of a conversation, and a *paralinguistic* (i.e., implicit) channel responsible for providing clues about the emotional state of the speaker and the implicit meaning of the message.

Information conveyed by the paralinguistic channel, in particular prosody, is useful for many research fields where the study of the rhythmic and intonational properties of speech is required (Leung et al., 2010). The ability to guess the emotional state of the speaker, as well as her/his communication style, are particularly interesting for Conversational Agents, as could allow them to select the more appropriate reaction to the user's requests, making the conversation more natural and thus improving the effectiveness of the system (Pleva et al., 2011; Moridis and Economides, 2012). Moreover, being able to extract paralinguistic information is interesting in clinic application, where psychological profiles of subjects and the clinical relationships they establish with doctors could be created. Finally, in forensic applications, paralinguistic information could be useful for observing how defendants, witnesses, and victims behave under interrogation.

Our contribution lies in the latter research field; in particular, we explore techniques for emotion and communication style recognition. In this paper we present an original model, a prototype (PrEmA - Prosodic Emotion Analyzer), and the results we obtained.

The paper is structured as follow. In Section 2 we provide a brief introduction about the relationship among voice, emotions, and communication styles; in Section 3 we present some research projects about emotion recognition; in Section 4 we introduce our model; in Section 5 we illustrate the experiments we conducted, and discuss the results we gathered; in Section 6 we introduce PrEmA, the prototype we built; finally, in Section 7 we draw some conclusions and outline our future research directions.

Sbattella L., Colombo L., Rinaldi C., Tedesco R., Matteucci M. and Trivilini A..
Extracting Emotions and Communication Styles from Vocal Signals.
DOI: 10.5220/0004699301830195
In *Proceedings of the International Conference on Physiological Computing Systems* (PhyCS-2014), pages 183-195
ISBN: 978-989-758-006-2

2 BACKGROUND

2.1 The Prosodic Elements

Several studies investigate the issue of characterizing human behaviors through vocal expressions; such studies rely on prosodic elements that transmit essential information about the speaker's attitude, emotion, intention, context, gender, age, and physical condition (Caldognetto and Poggi, 2004; Tesser et al., 2004; Asawa et al., 2012).

Intonation makes spoken language very different from written language. In written language, white spaces and punctuation are used to separate words, sentences and phrases, inducing a particular "rhythm" to the sentences. Punctuation also contributes to specify the meaning to the whole sentence, stressing words, creating emphasis on certain parts of the sentence, etc. In spoken language, a similar task is done by means of *prosody*—changes in speech rate, duration of syllables, intonation, loudness, etc.

Such so-called *suprasegmental* characteristics play an important role in the process of utterance understanding; they are key elements in expressing the *intention* of a message (interrogative, affirmative, etc.) and its *style* (aggressive, assertive, etc.) In this work we focused on the following prosodic characteristics (Pinker and Prince, 1994): intonation, loudness, duration, pauses, timbre, and rhythm.

***Intonation*(or *tonal Variation*, or *Melodic Contour*)** is the most important prosodic effects, and determines the evolution of speech melody. Intonation is tightly related to the illocutionary force of the utterance (e.g., assertion, direction, commission, expression, or declaration). For example, in Italian intonation is the sole way to distinguish among requests (by raising the intonation of the final part of the sentence), assertions (ascending intonation at the beginning of the sentence, then descending intonation in the final part), and commands (descending intonation); thus, it is possible to distinguish the question "vieni domani?" (are you coming tomorrow?) from the assertion "vieni domani" (you are coming tomorrow) or the imperative "vieni domani!" (come tomorrow!).

Moreover, intonation provides clues on the distribution of information in the utterance. In other words, it helps in emphasizing new or important facts the speaker is introducing in the discourse (for example, in Italian, by means of a peak in the intonation contour). Thus, intonation takes part in clarifying the syntactic structure of the utterance.

Finally, and most important for our work, intonation is also related to emotions; for example, the melodic contour of anger is rough and full of sudden variations on accented syllables, while joy exhibits a smooth, rounded, and slow-varying intonation. Intonation also conveys the attitude of the speaker, leading the hearer to grasp nuances of meaning, like irony, kindness, impatience, etc.

Loudness is another important prosodic feature, and is directly related to the voice loudness. Loudness can emphasize differences in terms of meaning—an increase of loudness, for example, can be related to anger.

***Duration* (or *Speech Rate*)** indicates the length of phonetic segments. Duration can transmit a wide range of meanings, such as speaker's emotions; in general, emotional states that imply psychophysiological activation (like fear, anger, and joy) are correlated to short durations and high speech rate (Bonvino, 2000), while sadness is typically related to slow speech. Duration also correlates with the speaker's attitudes (it gives clues about courtesy, impatience, or insecurity of the speaker), as well as types of discourse (a homily will have slower speech rate than, for example, a sport running commentary).

Pauses allow the speaker to take breath, but can also be used to emphasize parts of the utterance, by inserting breaks in the intonation contour; from this point of view, pauses correspond to punctuation we add in written language. Pauses, however, are much more general and can convey a larger variety of nuances than punctuation.

Timbre –such as falsetto, whisper, hoarse voice, quavering voice– often provide information about the emotional state and health of the speaker (for example, a speaker feeling insecure is easily associated with quavering voice). Timbre also depends on the mount of noise affecting the vocal emission.

Rhythm is a complex prosodic element, emerging from joint action of several factors, in particular intonation, loudness, and duration. It is an intrinsic and unique attribute of each language.

In the following, we present some studies that try to model the relationship among prosody, emotions, and communication styles.

2.2 Speech and Emotions

Emotion is a complex construct and represents a component of how we react to external stimuli (Scherer,

2005). In emotions we can distinguish:

- A neurophysiological component of activation (arousal).
- A cognitive component, through which an individual evaluates the situation-stimulus in relation to her/his needs.
- A motoric component, which aims at transforming intentions in actions.
- An expressive component, through which an individual expresses her/his intentions in relation to her/his level of social interaction.
- A subjective component, which is related to the experience of the individual.

The emotional expression is not only based on linguistic events, but also on paralinguistic events, which can be acoustic (such as screams or particular vocal inflections), visual (such as facial expressions or gestures), tactile (for example, a caress), gustatory, olfactory, and motoric (Balconi and Carrera, 2005; Planet and Iriondo, 2012). In particular, the contribution of non-verbal channels on the communication process is huge; according to (Mehrabian, 1972) the linguistic, paralinguistic, and motoric channels, constitutes, respectively, 7%, 38%, and 55% of the communication process. In this work, we focused on the acoustic paralinguistic channel.

According to (Stern, 1985), emotions can be divided in: *vital affects* (floating, vanishing, spending, exploding, increasing, decreasing, bloated, exhausted, etc.) and *categorical affects* (happiness, sadness, anger, fear, disgust, surprise, interest, shame). The former are very difficult to define and recognize, while the latter can be more easily treated. Thus, in this work we focused on categorical affects.

Finally, emotions have two components—an *hedonic tone*, which refers to the degree of pleasure, or displeasure, connected to the emotion; and an *activation*, which refers to the intensity of the physiological activation (Mandler, 1984). In this work we relied on the latter component, which is easier to measure.

2.2.1 Classifying Emotions

Several well-known theories for classifying emotions have been proposed. In (Russell and Snodgrass, 1987) authors consider a huge number of characteristics about emotions, identifying two primary axes: pleasantness / unpleasantness and arousal / inhibition.

In (Izard, 1971) the author lists 10 primary emotions: sadness, joy, surprise, sadness, anger, disgust, contempt, fear, shame, guilt; in (Tomkins, 1982) the latest one is eliminated; in (Ekman et al., 1994) a more restrictive classification (happiness, surprise, fear, sadness, anger, disgust) is proposed.

In particular, Ekman distinguishes between *primary emotions*, quickly activated and difficult to control (for example, anger, fear, disgust, happiness, sadness, surprise), and *secondary emotions*, which undergo social control and cognitive filtering (for example, shame, jealousy, pride). In this work we focused on primary emotions.

2.2.2 Mapping Speech and Emotions

As stated before, voice is considered a very reliable indicator of emotional states. The relationship between voice and emotion is based on the assumption that the physiological responses typical of an emotional state, such as the modification of breathing, phonation and articulation of sounds, produce detectable changes in the acoustic indexes associated to the production of speech.

Several theories have been developed in an effort to find a correlation among speech characteristics and emotions. For example, for Italian (Anolli and Ciceri, 1997):

- Fear is expressed as a subtle, tense, and tight tone.
- Sadness is communicated using a low tone, with the presence of long pauses and slow speech rate.
- Joy is expressed with a very sharp tone and with a progressive intonation profile, with increasing loudness and, sometimes, with an acceleration in speech rate.

In (Anolli, 2002) it is suggested that active emotions produce faster speech, with higher frequencies and wider loudness range, while the low-activation emotions are associated with slow voice and low frequencies.

In (Juslin, 1997) the author proposes a detailed study of the relationship between emotion an prosodic characteristics. His approach is based on time, loudness, spectrum, attack, articulation, and differences in duration (Juslin, 1998). Table 1 shows such prosodic characterization, for the four primary emotions; our work started from such clues, trying to derive measurable acoustic features.

Relying on the aforementioned works, we decided to focus on the following emotions: *joy*, *fear*, *anger*, *sadness*, and *neutral*.

2.3 Speech and Communication Styles

The process of communication has been studied from many points of view. Communication not only conveys information and expresses emotions, it is also

characterized by a particular relational style (in other words, a *communication style*). Everyone has a relational style that, from time to time, may be more or less dominant or passive, sociable or withdrawn, aggressive or friendly, welcoming or rejecting.

2.3.1 Classifying Communication Styles

We chose to rely on the following simple classification and description that includes three communication styles (Michel, 2008):

- Passive
- Assertive
- Aggressive

Passive communication imply not expressing honest feelings, thoughts and beliefs. Therefore, allowing others to violate your rights; expressing thoughts and feelings in an apologetic, self-effacing way, so that others easily disregard them; sometimes showing a subtle lack of respect for the other person's ability to take disappointments, shoulder some responsibility, or handle their own problems.

Persons with aggressive communication style stand up for their personal rights and express their thoughts, feelings and beliefs in a way which is usually inappropriate and always violates the rights of the other person. They tend to maintain their superiority by putting others down. When threatened, they tend to attack.

Finally, assertive communication is a way of communicating feelings, thoughts, and beliefs in an open, honest manner without violating the rights of others. It is an alternative to being aggressive where we abuse other people's rights, and passive where we abuse our own rights.

It is useful to learn the distinction among the aggressive, passive, and assertive communication behaviors, because such psychological characteristics provides clues on the prosodic parameters we can expect.

Table 1: Prosodic characterization of emotions.

Emotion	Prosodic feature
Joy	- quick meters - moderate duration variations - high average sound level - tendency to tighten up the contrasts between long and short words - articulation predominantly detached - quick attacks - brilliant tone - slight or missing vibrato - slightly rising intonation
Sadness	- slow meter - relatively large variations in duration - low noise level - tendency to attenuate the contrasts between long and short words - articulation linked - soft attacks - slow and wide vibrato - final delaying - soft tone - intonation (at times) slightly declining
Anger	- quick meters - high noise level - relatively sharp contrasts between long and short words - articulation mostly not linked - very dry attacks - sharply stamp - distorted notes
Fear	- quick meters - high noise level - relatively sharp contrasts between long and short words - articulation mostly not linked - very dry attacks - sharply stamp - distorted notes

2.3.2 Mapping Speech and Communication Styles

Starting from the aforementioned characteristics of communication styles, considering the prosodic clues provided in (Michel, 2008), and taking into account other works (Hirshberg and Avesani, 2000; Shriberg et al., 2000; Shriberg and Stolcke, 2001; Hastie et al., 2001; Hirst, 2001), we came out with the prosodic characterization showed in Table 2.

2.4 Acoustic Features

As we discussed above, characterizing emotional states and communication styles associated to a vocal signal implies measuring some acoustic features, which, in turn, are derived from physiological reactions. Table 1 and Table 2 provide some clues about how to relate such physiological reactions to prosodic characteristics, but we need to define a set of measurable acoustic features.

2.4.1 Acoustic Features for Emotions

We started from the most studied acoustic features (Murray and Arnott, 1995; McGilloway et al., 2000; Cowie et al., 2001; Wang and Li, 2012).

Pitch measures the intonation, and is represented by the fundamental harmonic (F0); it tends to increase for anger, joy, and fear; it decreases for sadness. Pitch

Table 2: Prosodic characterization of communication styles.

Communication style	Prosodic feature
Passive	- flickering - voice often dull and monotonous - tone may be sing-song or whining - low Volume - hesitant, filled with pauses - slow-fast or fast-slow - frequent throat clearing
Aggressive	- very firm voice - often abrupt, clipped - often fast - tone sarcastic, cold, harsh - grinding - fluent, without hesitations - voice can be strident, often shouting, rising at end
Assertive	- firm, relaxed voice - steady even pace - tone is middle range, rich and warm - not over-loud or quiet - fluent, few hesitation

tends to be more variable for anger and joy.

Intensity represents the amplitude of the vocal signal, and measures the loudness; intensity tends to increase for anger and joy, decrease for sadness, and stay constant for fear.

Time measures duration and pauses, as voiced and unvoiced segments. High speech rate is associated to anger, joy, and fear while low speech rate is associated to sadness. Irregular speech rate is often associated with anger and sadness. Time is also an important parameter for distinguishing articulation breaks (speaker's breathing) from *unvoiced* segments. The unvoiced segments represent *silences*—parts of the signal where the information of the pitch and/or intensity are below a certain threshold.

Voice Quality measures the timbre and is related to variations of the voice spectrum, as to the signal-noise ratio. In particular:

- Changes in amplitude of the waveform between successive cycles (called *shimmer*).
- Changes in the frequency of the waveform between successive cycles (called *jitter*).
- Hammarberg's index, which covers the difference between the energy in the 0-2000 Hz and 2000-5000 Hz bands.
- The harmonic/noise ratio (HNR) between the energy of the harmonic part of the signal and the remaining part of the signal; see (Hammarberg et al., 1980; Banse and Sherer, 1996; Gobl and Chasaide, 2000).

High values of shimmer and jitter characterize, for example, disgust and sadness, while fear and joy are distinguished by different values of the Hammarberg's index.

2.4.2 Acoustic Features for Communication Styles

Starting from clues provided by Table 2, we decided to rely on the same acoustic features we used for the emotion recognition (pitch, intensity, time, and voice quality). But, in order to recognize complex prosodic variations that particularly affects communication style, we reviewed the literature and found that research mostly focuses on the variations of tonal accents within a sentence and at a level of prominent syllables (Avesani et al., 2003; D'Anna and Petrillo, 2001; Delmonte, 2000). Thus, we decided to add two more elements to our acoustic feature set:

- Contour of the pitch curve
- Contour of the intensity curve

In Section 4.2 we will show how we measured the feature set we defined for emotions and communication style.

3 RELATED WORK

Several approaches exist in literature for the task of emotion recognition, based on classifiers like Support Vector Machines (SVM), decision trees, Neural Networks (NN), etc. In the following, we present some of such approaches.

The system described in (Koolagudi et al., 2011) made use of SVM for classifying emotions expressed by a group of professional speakers. The authors underlined that, for extreme emotions (anger, happiness and fear), the most useful information was contained in the first words of the sentence, while last words were more discriminative in case of neutral emotion. The recognition Precision [1] of the system, on average, using prosodic parameters and considering only the beginning words, was around 36%.

[1]Notice that the performance index provided in this section are indicative and cannot be compared each other, since each system used its own vocal dataset.

The approach described in (Borchert and Diisterhoft, 2005) used SVM, too, applying it to the German language. In particular, this project developed a prototype for analyzing the mood of customers in call centers. This research showed that pitch and intensity were the most important features for the emotional speech, while features on spectral energy distribution were the most important voice quality features. Recognition Precision they obtained was, on average, around 70%.

Another approach leveraged the Alternating Decision Trees (ADTree), for the analysis of humorous spoken conversations from a classic comedy TV show (Purandare and Litman, 2006); speaker turns were classified as humorous or non-humorous. They used a combination of prosodic (e.g., pitch, energy, duration, silences, etc.) and non-prosodic features (e.g., words, turn length, etc.) Authors discovered that the best set of features was related to the gender of the speaker. Their classifier obtained Accuracies of 64.63% for males and 64.8% for females.

The project described in (Shi and Song, 2010) made use of NN. The project used two databases of Chinese utterances. One was composed of speech recorded by non-professional speakers, while the other was composed of TV recordings. They used Mel-Frequency Cepstral Coefficients for analyzing the utterances, considering six speech emotions: angry, happiness, sadness, and surprised. They obtained the following Precisions: angry 66%, happiness 57.8%, sadness 85.1%, and surprised 58.7%.

The approaches described in (Lee and Narayanan, 2005) used a combination of three different sources of information: acoustic, lexical, and discourse. They proposed a case study for detecting negative and non-negative emotions using spoken language coming from a call center application. In particular, the samples were obtained from real users involved in spoken dialog with an automatic agent over the telephone. In order to capture the emotional features at the lexical level, they introduced a new concept named "emotional salience"—an emotionally salient word, with respect to a category, tends to appear more often in that category than in other categories. For the acoustic analysis they compared a K-Nearest Neighborhood classifier and a Linear Discriminant Classifier. The results of the project demonstrated that the best performance was obtained when acoustic and language features were combined. The best performing results of this project, in terms of classification errors, were 10.65% for males and 7.95% for females.

Finally, in (López-de Ipiña et al., 2013) authors focuses on "emotional temperature" (ET) as a biomarker for early Alzheimer disease detection. They leverages non linear features, such as the Fractal Dimension, and rely on a SVM for classifying ET of voice frames as pathological or non-pathological. They claim an Accuracy of 90.7% to 97.7%.

Our project is based on a classifier that leverages the Linear Discriminant Analysis (LDA) (McLachlan, 2004); such a model is simpler than SVM and NN, and easier to train. Moreover, with respect to approaches making use of textual features, our model is considerably simpler. Nevertheless, our approach provides good results (see Section 5).

Finally, we didn't find any system able to classify communication styles so, to our knowledge, this feature provided by our system is novel.

4 THE MODEL

For each voiced segment, two set of features –one for recognizing emotions and one for communication style– were calculated; then, by means of two LDA-based classifiers, such segments were associated with emotion and communication style.

LDA-based classifier provided a good trade-off between performance and classification correctness. LDA projects vectors of features, which represents the samples to analyze, to a smaller space. The method maximizes the ratio of between-class variance to the within-class variance, permitting to maximize class separability. More formally, LDA finds the eigenvectors $\vec{\phi}_i$ that solve:

$$\mathbf{B}\vec{\phi}_i - \lambda \mathbf{W}\vec{\phi}_i = 0 \tag{1}$$

where $\mathbf{B}$ is the between-class scatter matrix and $\mathbf{W}$ is the within-class scatter matrix. Once a sample $\vec{x}_j$ is projected on the new space provided by the eigenvectors, the class $\hat{k}$ corresponding to the projection $\vec{y}_j$ is chosen according to (Boersma and Weenink, 2013):

$$\hat{k} = \underset{k}{\operatorname{argmax}}\, p(k|\vec{y}_j) = \underset{k}{\operatorname{argmax}} -d_k^2(\vec{y}_j) \tag{2}$$

where $d_k^2(\cdot)$ is the generalized squared distance function:

$$d_k^2(\vec{y}) = (\vec{y}-\vec{\mu}_j)^T \Sigma_k^{-1} (\vec{y}-\vec{\mu}_j) + \frac{\ln|\Sigma_k|}{2} - \ln p(k) \tag{3}$$

where Σ_k is the covariance matrix for the class k and $p(k)$ is the a-priori probability of the class k:

$$p(k) = \frac{n_k}{\sum_{i=1}^{K} n_i} \tag{4}$$

where n_k is the number of samples belonging to the class k, and K is the number of classes.

4.1 Creating a Corpus

Our model was trained and tested on a corpus of sentences, labeled with the five basic emotions and the three communication styles we introduced. We collected 900 sentences, uttered by six Italian professional speakers, asking them to simulate emotions and communication styles. This way, we obtained good samples, showing clear emotions and expressing the desired communication styles.

4.2 Measuring and Selecting Acoustic Features

Figure 1 shows the activities that lead to the calculation of the acoustic features: Preprocessing, segmentation, and feature extraction. The result is the dataset we used for training and testing the classifiers.

In the following, the aforementioned phases are presented. The values shown for the various parameters needed by the voice-processing routines, have been chosen experimentally (see Section 4.3.2 for details on how the values of such parameters were selected; see Section 6 for details on Praat, the voice-processing tool we adopted).

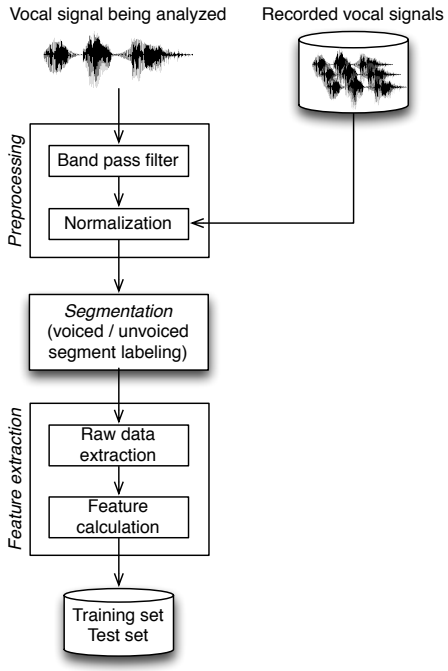


Figure 1: The feature calculation process.

4.2.1 Preprocessing and Segmentation

We used a Hann band filter, for removing useless harmonics (F_{Lo}=100Hz, F_{Hi}=6kHz, and smoothing w=100Hz). Then, we normalized the intensity of different audio files, so that the average intensity of different recordings was uniform and matched a predefined setpoint. Finally, we divided the audio signal into *segments*; in particular we divided *voiced segments*, where the average, normalized intensity was above the threshold $I_{voicing}$=0.45, and *silenced segments*, where the average, normalized intensity was below the threshold $I_{silence}$=0.03. Segments having average, normalized intensity between the two thresholds were not considered [2].

4.2.2 Feature Calculation

For features related to Pitch, we used the following range F_{floor}=75Hz, $F_{ceiling}$=600Hz (such values are well suited for male voices, as we used male subjects for our experiments).

Among features related with Time, *articulation ratio* refers to the amount of time taken by voiced segments, excluding articulation breaks, divided by the total recording time; an *articulation break* is a pause –in a voiced segment– longer than a given threshold (we used the threshold T_{break}=0.25s), and is used to capture the speaker's breathing. The *speech ratio*, instead, is the percentage of voiced segments over the total recording time. These two parameters are very similar for short utterances, because articulation breaks are negligible; for long utterances, however, these parameters definitely differ, revealing that articulation breaks are an intrinsic property of the speaker. Finally, *unvoiced frame ratio* is the total time of unvoiced frames, divided by the recording total time

The speech signal, even if produced with maximum stationarity, contains variations of F0 and intensity (Hammarberg et al., 1980); such variations represents the perceived voice quality. The random changes in the short term (micro disturbances) of F0 are defined as *jitter*, while the variations of the amplitude are known as *shimmer*. The Harmonic-to-Noise ratio (HNR) value is the "degree of hoarseness" of the signal—the extent to which noise replaces the harmonic structure in the spectrogram (Boersma, 1993).

Finally, the following features are meant to represent Pitch and Intensity contours:

- Pitch Contour
  - Number of peaks per second. The number of maxima in the pitch contour, within a voiced segment, divided by the duration of the segment.
  - Average and variance of peak values.

[2] Such segments were considered too loud for being clear silences, but too quiet for providing a clear voiced signal.

  - Average gradient. The average gradient between two consecutive sampling points in the pitch curve.
  - Variance of gradients. The variance of such pitch gradients.
- Intensity Contour
  - Number of peaks per second. The number of maxima in the intensity curve, within a voiced segment, divided by the duration of the segment.
  - Mean and variance of peak values.
  - Variance of peak values.
  - Average gradient. The average gradient between two consecutive sampling points in the intensity curve.
  - Variance of gradients. The variance of such intensity gradients.

Table 3 and Table 4 summarized the acoustic features we measured, for emotions and communication styles, respectively.

Table 3: Measured acoustic features for emotions.

Features	Characteristics
Pitch (F0)	Average [Hz] Standard deviation [Hz] Maximum [Hz] Minimum [Hz] 25th quantile [Hz] 75th quantile [Hz] Median [Hz]
Intensity	Average [dB] Standard deviation [dB] Maximum [dB] Minimum [dB] Median [dB]
Time	Unvoiced frame ratio [%] Articulation break ratio [%] Articulation ratio [%] Speech ratio [%]
Voice quality	Jitter [%] Shimmer [%] HNR [dB]

The features we defined underwent a selection process, aiming at discarding highly correlated measurements, in order to obtain the minimum set of features. In particular, we used the ANOVA and the LSD tests.

The ANOVA analysis for features related to emotions (assuming 0.01 as significance threshold) found all the features to be significant, except Average_Intensity. For Average_intensity we leveraged the Fischer's LSD test, which showed that Average_Intensity was not useful for discriminating Joy from Neutral and Sadness, Neutral from Sadness, and Fear from Anger. Nevertheless Average_Intensity was retained, as LSD proved it useful for discriminating Sadness, Joy, and Neutral from Anger and Fear.

Table 4: Measured acoustic features for communication style (features in italics have been removed).

Features	Characteristics
Pitch (F0)	Average [Hz] Standard deviation [Hz] Maximum [Hz] Minimum [Hz] 10th quantile [Hz] 90th quantile [Hz] Median [Hz]
Pitch contour	Peaks per second [#peaks/s] Average peaks height [Hz] Variance of peak heights [Hz] Average peak gradient [Hz/s] Variance of peak gradients [Hz/s]
Intensity	*Average [dB]* Standard deviation [dB] Maximum [dB] Minimum [dB] 10th quantile [dB] *90th quantile [dB]* Median [dB]
Intensity contour	Peaks per second [#peaks/s] *Average peak height [dB]* *Variance of peaks heights [dB]* *Average peak gradients [dB/s]* *Variance of peak gradients [dB/s]*
Time	*Unvoiced frame ratio [%]* Articulation break ratio [%] Articulation ratio [%] Speech ratio [%]
Voice quality	Jitter [%] Shimmer [%] NHR [dB]

The ANOVA analysis for features related to communication style (assuming 0.01 as significance threshold) found eight potentially useless features: Average_Intensity, 90_th_quantile, Unvoiced_frame_ratio, Peaks_per_second, Average_peak_gradient, Standard_deviation_peaks_gradient, Median_intensity, and Standard_deviation_intensity. For such features we performed the LSD test, which showed that Peaks_per_second was not able to discriminate Aggressive vs Assertive, but was useful for discriminating all others communication styles and thus we decided to retain it. The others seven features were dropped as LSD showed that they were not useful for discriminating communication styles.

After this selection phase, the set of features for the emotion recognition task remained unchanged,

while the set of features for the communication-style recognition task was reduced (in Table 4, text in italics indicates removed features).

4.3 Training

4.3.1 The Vocal Dataset

For the creation of the vocal corpus we examined the public vocal databases available for the Italian language (EUROM0, EUROM1 and AIDA), public audiobooks, and different resources provided by professional actors. After a detailed evaluation of available resources, we realized that they were not suitable for our study, due to the scarcity of sequences where emotion an communication style were unambiguously expressed.

We therefore opted for the development of our own datasets, composed of:

- A series of sentences, with different emotional intentions
- A series of monologues, with different communication styles

We carefully selected –taking into account the work presented in (Canepari, 1985)– 10 sentences for each emotion, expressing strong and clear emotional states. This way, it was easier for the actor to communicate the desired emotional state, because the meaning of the sentence already contained the emotional intention. With the same approach we selected 3 monologues (about ten to fifteen rows long, each)—they were chosen to help the actor in identifying himself with the desired communication style.

For example, to represent the passive style we chose some monologues by Woody Allen; to represent the aggressive style, we chose "The night before the trial" by Anton Chekhov; and to represent assertive style, we used the book "Redesigning the company" by Richard Normann.

We selected six male actors; each one was recorded independently and individually, in order to avoid mutual conditioning. In addition, each actor received the texts in advance, in order to review them and practice before the registration.

4.3.2 The Learning Process

The first step of the learning process was to select the parameters needed by the voice-processing routines. Using the whole vocal dataset we trained several classifiers, varying the parameters, and selected the best combination according to the performance indexes we defined (see Section 5). We did it for both the emotion recognition classifier and the communication-style recognition classifier, obtaining two parameter sets.

Once the parameter sets were defined, a subset of the vocal dataset –the training dataset– was used to train the two classifiers. In particular, for the emotional dataset –containing 900 voiced segments– and the communication-style dataset –containing 54 paragraphs– we defined a training dataset containing 90% of the initial dataset, and an evaluation dataset containing the remaining 10%.

Then, we trained the two classifiers on the training dataset. Such process was repeated 10 times, with different training set/test set subdivisions.

Table 5: Confusion matrix for emotions (%).

	Predicted emotions				
	Joy	Neutral	Fear	Anger	Sadness
Joy	**63.81**	0.00	18.35	11.79	6.05
Neutral	3.47	**77.51**	2.14	1.79	15.09
Fear	33.75	0.00	**58.35**	6.65	1.25
Anger	10.24	1.16	8.16	**77.28**	3.16
Sadness	5.14	14.44	0.28	0.81	**79.33**

5 EVALUATION AND DISCUSSION

During the evaluation phase, the 10 pairs of LDA-based classifiers we trained (10 for emotions and 10 for communication styles) tagged each voiced segment in the evaluation dataset with an emotion and an communication style. Then performance metrics were calculated for each classifier; finally, average performance metrics were calculated (see Section 6 for details on Praat, the voice-processing tool we adopted).

5.1 Emotions

The validation dataset consists of 18 voiced segments chosen at random for each of the five emotions, for a total of 90 voiced segments (10% of the whole emotion dataset).

The average performance indexes of the 10 trained classifiers, are shown in Table 5 and Table 6

Precision and F-measure are good for Neutral, Anger, and Sadness, while Fear and Joy are more problematic (especially Joy, which has the worst value). The issue is confirmed by the confusion matrix of Table 5, which shows that Joy phrases were

Table 6: Precision, Recall, and F-measure for emotions (%).

	Joy	Neutral	Fear	Anger	Sadness
Pr	56.03	75.00	64.84	80.36	80.28
Re	63.53	76.36	58.42	77.10	79.39
F_1	59.54	75.68	61.46	78.70	79.83

Table 7: Error rates for emotions.

	Joy	Neutral	Fear	Anger	Sadness
F_p	164	56	96	65	70
F_n	120	52	126	79	74
T_e **(%)**	**18.25**	**6.94**	**14.27**	**9.25**	**9.25**

Table 8: Average *Pr*, *Re*, F_1, and *Ac*, for emotions (%).

Avg *Pr*	71.44
Avg *Re*	71.06
Avg F_1	71.16
Ac	71.27

tagged as Fear 33% of the time, lowering the Precision of both. Recall is good for all the emotions and also for Joy, which exhibits the better value. The average values for Precision, Recall, and F-measure are about 71%; Accuracy exhibits a similar value.

The *K* value, the agreement between the classifier and the dataset, is *K*=0.63541, meaning a good agreement was found.

Finally, for each class, we calculated false positives F_p (number of voiced segments belonging to another class, incorrectly tagged in the class), false negatives F_n (number of voiced segments belonging to this class, incorrectly classified in another class), and thus the error rate T_e (see Table 7).

Joy and Fear exhibit the highest errors, as the classifier often confused them. We argue this result is due to the highs degree of arousal that characterize both Joy and Fear.

5.2 Communication Styles

The validation data set consists of 2 randomly chosen paragraphs, for each of the three communication styles, for a total of 6 paragraphs, which corresponds to 10% of the communication-style dataset.

The average performance indexes of the 10 trained models, are shown in Table 9 and Table 10.

Precision, Recall, and F-measure indicate very good performances for Aggressive and Passive communication styles; acceptable but much smaller values are obtained for Assertive sentences, as they are often tagged as Aggressive (24.26% of the time, as shown in the confusion matrix). The average values for Precision, Recall, and F-measure are about 86%; Accuracy exhibits a similar value.

Table 9: Confusion matrix for communication styles (%).

	Predicted communication styles		
	Aggressive	Assertive	Passive
Aggressive	**99.30**	0.70	0.00
Assertive	24.26	**62.68**	13.06
Passive	7.08	10.30	**82.62**

Table 10: Precision, Recall, and F-measure for communication styles (%).

	Aggressive	Assertive	Passive
Pr	85.55	68.32	93.61
Re	99.33	60.53	83.25
F_1	91.93	64.19	88.13

Table 11: Error rate for communication style.

	Aggressive	Assertive	Passive
F_p	100	64	36
F_n	4	90	106
T_e **(%)**	**7.14**	**10.57**	**9.75**

Table 12: Average *Pr*, *Re*, F_1, and *Ac*, for communication-style (%).

Avg *Pr*	86.10
Avg *Re*	85.87
Avg F_1	85.61
Ac	86.00

The *K* value, the agreement between the classifier and the dataset, is *K*=0.777214, meaning that a good agreement was found.

Finally, Table 11 shows error rates for each class.

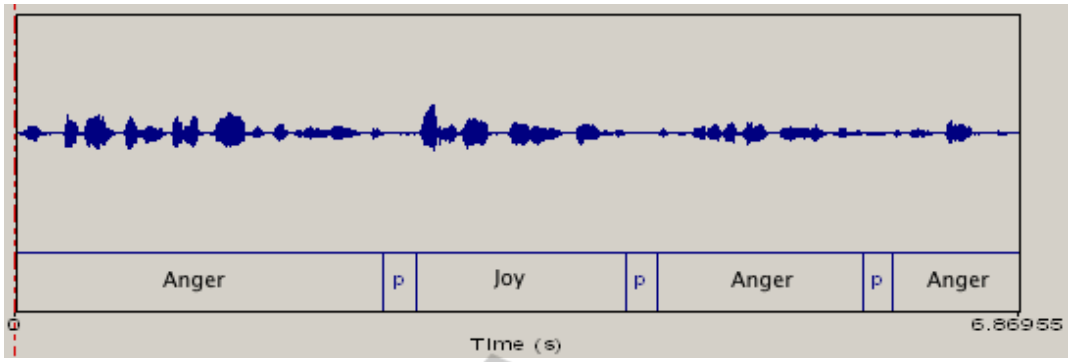
As expected, Assertive exhibits the highest error (10.57%), while the best result is achieved by the recognition of Aggressive, with an error rate of 7.14%. Analyzing the F_p and F_n values we noted that only Aggressive had $F_n > F_p$, which means that the classifier tended to mistakenly associate such a class to segments where it was not appropriate.

6 THE PROTOTYPE

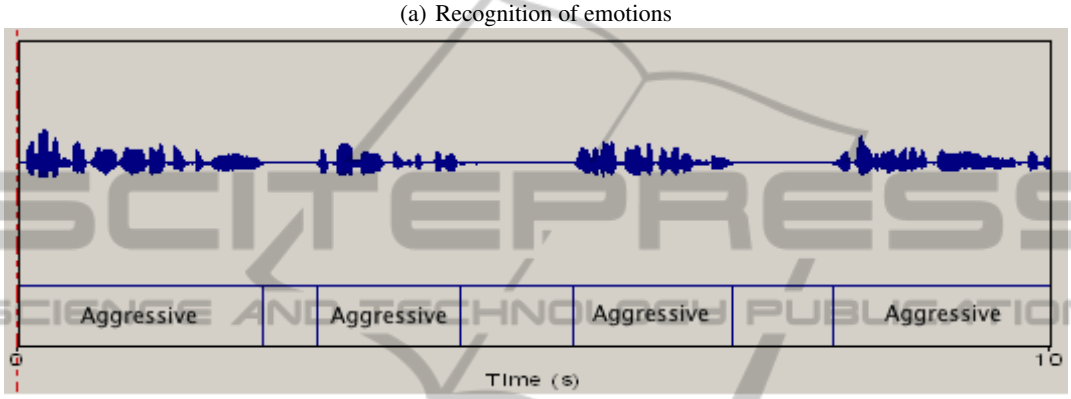
The application architecture is composed of five modules (see Figure 2): GUI, Feature Extraction, Emotion Recognition, Communication-style Recognition, and Praat.

Praat (Boersma, 2001) is a well-known open-source tool for speech analysis [3]; it provides several functionalities for the analysis of vocal signals as well as a statistical module (containing the LDA-based classifier we described in Section 4). The script-

[3] http://www.fon.hum.uva.nl/praat/



(a) Recognition of emotions



(b) Recognition of communication styles

Figure 3: Recognition of emotions and communication styles.

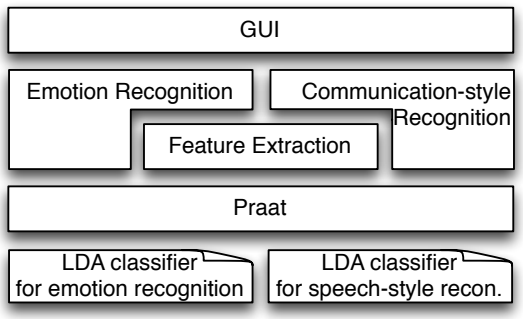


Figure 2: The PrEmA architecture overview.

ing functionalities provided by Praat permitted us to quickly implement our prototype.

The GUI module permits to choose the audio file to analyze, and shows the results to the user; the Feature Extraction module performs the calculations presented in Section 4 (preprocessing, segmentation, and feature calculation); the Emotion Recognition and Communication-style Recognition modules rely on the two best-performing [4] models to classify the input file according to its emotional state and communication style. All the calculations are implemented by means of scripts that leverage functionalities provided by Praat.

Figure 3 shows two screenshots of the PrEmA application (translated in English) recognizing, respectively, emotions and communication styles. In particular, in Figure 3(a) the application analyzed a sentence of 6.87s expressing anger, divided in four segments (i.e., three silences where found); each segment was assigned with an emotion: Anger, Joy (mistakenly), Anger, Anger. Figure 3(b) shows the first 10s fragment of a 123.9s aggressive speech; the application found 40 segments and assigned them with a communication style (in the example, the segments where all classified as Aggressive).

7 CONCLUSIONS AND FUTURE WORK

We presented PrEmA, a tool able to recognize emotions and communication styles from vocal signals, providing clues about the state of the conversation. In particular, we consider communication-style recognition as our main contribution since it could provide

[4]From the 10 LDA-based classifiers generated for the emotion classification task, the one with better performance indexes was chosen as a final model; the same approach was followed for the communication-style classifier.

a potentially powerful mean for understanding user's needs, problems and desires.

The tool, written using the Praat scripting language, relies on two sets of prosodic features and two LDA-based classifiers. The experiments, performed on a custom corpus of tagged audio recordings, showed encouraging results: for classification of emotions, we obtained a value of about 71% for average *Pr*, average *Re*, average F_1, and *Ac*, with a K=0.64; for classification of communication styles, we obtained a value of about 86% for average *Pr*, average *Re*, average F_1, and *Ac*, with a K=0.78.

As a future work, we plan to test other classification approaches, such as HMM and CRF, experimenting them with a bigger corpus. Moreover, we plan to investigate text-based features provided by NLP tools, like POS taggers and parsers. Finally, the analysis will be enhanced according to the "musical behavior" methodology (Sbattella, 2006; Sbattella, 2013).

REFERENCES

Anolli, L. (2002). *Le emozioni*. Ed. Unicopoli.

Anolli, L. and Ciceri, R. (1997). *The voice of emotions*. Milano, Angeli.

Asawa, K., Verma, V., and Agrawal, A. (2012). Recognition of vocal emotions from acoustic profile. In *Proceedings of the International Conference on Advances in Computing, Communications and Informatics*.

Avesani, C., Cosi, P., Fauri, E., Gretter, R., Mana, N., Rocchi, S., Rossi, F., and Tesser, F. (2003). Definizione ed annotazione prosodica di un database di parlato-letto usando il formalismo ToBI. In *Proc. of Il Parlato Italiano*, Napoli, Italy.

Balconi, M. and Carrera, A. (2005). Il lessico emotivo nel decoding delle espressioni facciali. *ESE - Psychofenia - Salento University Publishing*.

Banse, R. and Sherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*.

Boersma, P. (1993). Accurate Short-Term Analysis of the Fundamental Frequency and the Harmonics-to-Noise Ratio of a Sampled Sound. *Institute of Phonetic Sciences, University of Amsterdam, Proceedings*, 17:97–110.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345.

Boersma, P. and Weenink, D. (2013). Manual of praat: doing phonetics by computer [computer program].

Bonvino, E. (2000). *Le strutture del linguaggio: unintroduzione alla fonologia*. Milano: La Nuova Italia.

Borchert, M. and Diisterhoft, A. (2005). Emotions in speech - experiments with prosody and quality features in speech for use in categorical and dimensional emotion recognition environments. *Natural Language Processing and Knowledge Engineering, IEEE*.

Caldognetto, E. M. and Poggi, I. (2004). Il parlato emotivo. aspetti cognitivi, linguistici e fonetici. In *Il parlato italiano. Atti del Convegno Nazionale, Napoli 13-15 febbraio 2003*.

Canepari, L. (1985). *LIntonazione Linguistica e paralinguistica*. Liguori Editore.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., and Fellenz, W. (2001). Emotion recognition in human-computer interaction. *Signal Processing Magazine, IEEE*.

D'Anna, L. and Petrillo, M. (2001). Apa: un prototipo di sistema automatico per lanalisi prosodica. In *Atti delle 11e giornate di studio del Gruppo di Fonetica Sperimentale*.

Delmonte, R. (2000). Speech communication. In *Speech Communication*.

Ekman, D., Ekman, P., and Davidson, R. (1994). *The Nature of Emotion: Fundamental Questions*. New York Oxford, Oxford University Press.

Gobl, C. and Chasaide, A. N. (2000). Testing affective correlates of voice quality through analysis and resynthesis. In *ISCA Workshop on Emotion and Speech*.

Hammarberg, B., Fritzell, B., Gauffin, J., Sundberg, J., and Wedin, L. (1980). Perceptual and acoustic correlates of voice qualities. *Acta Oto-laryngologica*, 90(1–6):441–451.

Hastie, H. W., Poesio, M., and Isard, S. (2001). Automatically predicting dialog structure using prosodic features. In *Speech Communication*.

Hirshberg, J. and Avesani, C. (2000). *Prosodic disambiguation in English and Italian, in Botinis*. Ed., Intonation, Kluwer.

Hirst, D. (2001). Automatic analysis of prosody for multilingual speech corpora. In *Improvements in Speech Synthesis*.

Izard, C. E. (1971). *The face of emotion*. Ed. Appleton Century Crofts.

Juslin, P. (1998). *A functionalist perspective on emotional communication in music performance*. Acta Universitatis Upsaliensis, 1st edition.

Juslin, P. N. (1997). Emotional communication in music performance: A functionalist perspective and some data. In *Music Perception*.

Koolagudi, S. G., Kumar, N., and Rao, K. S. (2011). Speech emotion recognition using segmental level prosodic analysis. *Devices and Communications (ICDeCom), IEEE*.

Lee, C. M. and Narayanan, S. (2005). Toward detecting emotions in spoken dialogs. *Transaction on Speech and Audio Processing, IEEE*.

Leung, C., Lee, T., Ma, B., and Li, H. (2010). Prosodic attribute model for spoken language identification. In *Acoustics, speech and signal processing. IEEE international conference (ICASSP 2010)*.

López-de Ipiña, K., Alonso, J.-B., Travieso, C. M., Solé-Casals, J., Egiraun, H., Faundez-Zanuy, M., Ezeiza, A., Barroso, N., Ecay-Torres, M., Martinez-Lage, P., and Lizardui, U. M. d. (2013). On the selection of

non-invasive methods based on speech analysis oriented to automatic alzheimer disease diagnosis. *Sensors*, 13(5):6730–6745.

Mandler, G. (1984). *Mind and Body: Psychology of Emotion and Stress*. New York: Norton.

McGilloway, S., Cowie, R., Cowie, E. D., Gielen, S., Westerdijk, M., and Stroeve, S. (2000). Approaching automatic recognition of emotion from voice: a rough benchmark. In *ISCA Workshop on Speech and Emotion*.

McLachlan, G. J. (2004). *Discriminant Analysis and Statistical Pattern Recognition*. Wiley.

Mehrabian, A. (1972). *Nonverbal communication*. Aldine-Atherton.

Michel, F. (2008). *Assert Yourself*. Centre for Clinical Interventions, Perth, Western Australia.

Moridis, C. N. and Economides, A. A. (2012). Affective learning: Empathetic agents with emotional facial and tone of voice expressions. *IEEE Transactions on Affective Computing*, 99(PrePrints).

Murray, E. and Arnott, J. L. (1995). Towards a simulation of emotion in synthetic speech: a review of the literature on human vocal emotion. *Journal of the Acoustical Society of America*.

Pinker, S. and Prince, A. (1994). Regular and irregular morphology and the psychological status of rules of grammar. In *The reality of linguistic rules*.

Planet, S. and Iriondo, I. (2012). Comparison between decision-level and feature-level fusion of acoustic and linguistic features for spontaneous emotion recognition. In *Information Systems and Technologies (CISTI)*.

Pleva, M., Ondas, S., Juhar, J., Cizmar, A., Papaj, J., and Dobos, L. (2011). Speech and mobile technologies for cognitive communication and information systems. In *Cognitive Infocommunications (CogInfoCom), 2011 2nd International Conference on*, pages 1 –5.

Purandare, A. and Litman, D. (2006). Humor: Prosody analysis and automatic recognition for F * R * I * E * N * D * S *. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*, Sydney, Australia.

Russell, J. A. and Snodgrass, J. (1987). *Emotion and the environment*. Handbook of Environmental Psychology.

Sbattella, L. (2006). *La Mente Orchestra. Elaborazione della risonanza e autismo*. Vita e pensiero.

Sbattella, L. (2013). *Ti penso, dunque suono. Costrutti cognitivi e relazionali del comportamento musicale: un modello di ricerca-azione*. Vita e pensiero.

Scherer, K. (2005). What are emotions? and how can they be measured? *Social Science Information*.

Shi, Y. and Song, W. (2010). Speech emotion recognition based on data mining technology. In *Sixth International Conference on Natural Computation*.

Shriberg, E. and Stolcke, A. (2001). Prosody modeling for automatic speech recognition and understanding. In *Proc. of ISCA Workshop on Prosody in Speech Recognition and Understanding*.

Shriberg, E., Stolcke, A., Hakkani-Tr, D., and Tr, G. (2000). *Prosody-based automatic segmentation of speech into sentences and topics*. Ed. Speech Communication.

Stern, D. (1985). *Il mondo interpersonale del bambino*. Bollati Boringhieri, 1st edition.

Tesser, F., Cosi, P., Orioli, C., and Tisato, G. (2004). Modelli prosodici emotivi per la sintesi dell'italiano. *ITC-IRST, ISTC-CNR*.

Tomkins, S. (1982). *Affect theory*. Approaches to emotion, Ed. Lawrence Erlbaum Associates.

Wang, C. and Li, Y. (2012). A study on the search of the most discriminative speech features in the speaker dependent speech emotion recognition. In *Parallel Architectures, algortihms and programming. International symposium (PAAP 2012)*.