Audio-based Emotion Recognition for Advanced Automatic Retrieval in Judicial Domain

F. Archetti[1,2], G. Arosio[1], E. Fersini[1], E. Messina[1]

[1] DISCO, Università degli Studi di Milano-Bicocca,
Viale Sarca, 336 - 20126 Milano, Italy
{archetti, arosio, fersini, messina}@disco.unimib.it
[2] Consorzio Milano Ricerche,
Via Cicognara 7 - 20129 Milano, Italy
archetti@milanoricerche.it

Abstract. Thanks to the recent progresses in judicial proceedings management, especially related to the introduction of audio/video recording systems, semantic retrieval has now become a realistic key challenge. In this context emotion recognition engine, through the analysis of vocal signature of actors involved in judicial proceedings, could provide useful annotations for semantic retrieval of multimedia clips. With respect to the generation of semantic emotional tag in judicial domain, two main contributions are given: (1) the construction of an Italian emotional database for Italian proceedings annotation; (2) the investigation of a hierarchical classification system, based on risk minimization method, able to recognize emotional states from vocal signatures. In order to estimate the degree of affection we compared the proposed classification method with the traditional ones, highlighting in terms of classification accuracy the improvements given by a hierarchical learning approach.

1 Introduction

The IT infrastructure introduced into judicial environments, with particular attention at audio/video recording systems into courtrooms, had a great impact related to the legal actor work's. All the recorded events that occur during a trial are available for subsequent consultation. However, despite the huge quantity of information expressed in multimedia form that are captured during trials, the current retrieval process of contents is based on manual consultation of the entire multimedia tracks or, in the best case, on an automatic retrieval service based on textual user queries with no possibility to search specific semantic concepts. Innovative features, that will impact the current consultation processes, are introduced by the JUMAS project: fusion of semantic annotations of different data streams to deliver a more effective automatic retrieval system. A synthetic representation of JUMAS components are depicted in figure 1. Consorzio Milano Ricerche and Milano-Bicocca University will address in JUMAS three main topics: (1) semantic annotation of the audio stream, (2) automatic template filling of judicial transcripts and (3) multimedia summarization of audio/video judicial

proceedings. In this paper we deal with the semantic annotation of audio signals that characterize each trial.

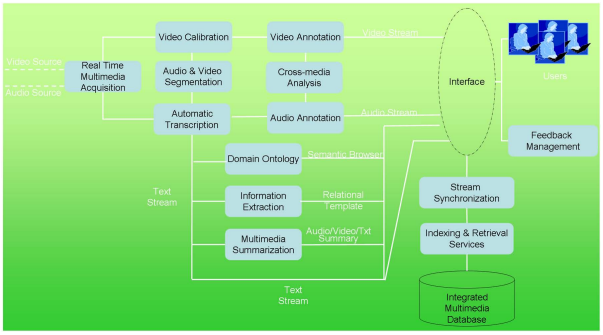


Fig. 1. JUMAS architecture

Emotional states associated to the actors involved in courtroom debates, represent one of the semantic concepts that can be extracted from multimedia sources, indexed and subsequently retrieved for consultation purposes. It is useful to stress the main difference between our method and the one at the base of Layered Voice Analysis (LVA) systems: while the main objective of LVA is to empower security officers and law enforcement agencies to discriminate between "normal stress" and stress induced by deception during investigative phases, in JUMAS the aim is to create semantic annotation of emotional state in order to allow "emotion-based" retrieval of multimedia judicial proceedings clips. Despite the progress in understanding the mechanisms of emotions in human speech from a psychological point of view, progress in the design and development of automatic emotion recognition systems for practical applications is still in its infancy, especially in judicial contexts. This limited progress is due to several reasons: (1) representation of vocal signal with a set of numerical features able to achieve reliable recognition; (2) identification of those emotional states that derive from a composition of other emotions (for example the "remorse" emotion is a combination of "sadness" and "disgust"); (3) presence of inter-speaker differences such as the variation in language and culture; (4) noisy environment; (5) interaction among speakers; (6) quality of the emotional database used for learning, and its likelihood with the real world uttered emotions. A general emotion recognition process can be described by four main phases: dataset construction, attribute extraction, feature selection/generation and inference model learning. The first phase deals with the collection of a corpus of voice signals uttered by different speakers and representative of several emotional states. When the database is created, the features extraction step is performed in order to map the vocal signals into descriptive attributes collected in a series of numerical vectors. Among this attributes through a feature selection/construction phase, a

feature set able to better discriminate emotional states is derived. This features are used in the final step to create a classification model able to infer emotional states of unlabelled speakers. With respect to these four main phases the literature can be classified accordingly. Concerning the dataset construction step, several benchmarks in different language have been collected. Among other we can find Serbian [5], German [2] and Polish [3] emotional corpus. Considering the attribute extraction phase, two of the most comprehensive studies ([1] and [10]) were aimed at discovering those attribute set that better correlates with respect to a given collection of emotional states. Their results highlighted that F_0 or spectral information have high impact in automatic emotion recognition systems. With respect to the feature selection step, there exists a great number of approaches aimed at identifying the most discriminative characteristics for a set of emotional states [6] [7] [8]. Concerning the final step related to the induction of inference models, able to recognize emotional states of unlabelled speaker, classification algorithms were widely investigated. The most extensive comparisons between several classification algorithms are reported in [9] and [10].

In this paper, we address the problem of finding the model that, with respect to courtroom debates characteristics, is able to produce the optimal recognition performance. The outline of the paper is the following. In section 2 we present two emotional corpus. A well-known benchmark for the German language is introduced, while a new benchmark is proposed for the Italian language. In Section 3 the extraction of vocal signature from uttered emotional sentences is described. In Section 4 traditional inference models and the proposed Multi-Layer Support Vector Machines approach, with their respective experimental results, are described. Finally, in Section 5 conclusions are derived.

2 Emotion Corpus

The performance of an automatic emotion recognition system strictly depends on the quality of the database used for inducing an inference model. Since an emotion recognition engine must be "trained" by a set of samples, i.e. needs to estimate model parameters through a set of emotionally labelled sentences, there are three ways of obtaining an emotional corpus:

1. recording by professional actors: the actors identify themselves in a specific situation before acting a given "emotional" sentence;
2. Wizard-of-Oz (WOZ): a system interacts with the actors and guide them into a specific emotional state that is subsequently recorded;
3. recording of real-word human emotions: the "emotional" sentences are gathered by recording real life situations.

In order to compare the performance of learning algorithms with the state of the art, we choose from the literature one of the most used emotional corpus known as *Berlin Database of Emotional Speech* or *Emo-DB*. This emotional corpus is composed by a set of wave files (531 samples) that represent different emotional

states: *neutral, anger, fear, joy, sadness, disgust and boredom*. For a more detailed description refers to [2]. A further benchmark, built at the University of Milano-Bicocca, is presented in the next subsection.

2.1 Italian Emotional DB

As pointed out in section 1, emotion recognition can be strongly influenced by several factors, and in particular by language and culture. For this reason, we decided that it would be useful to adopt an Italian corpus in order to investigate Italian emotional behaviors. Since at the time of writing there is no Italian benchmark, we decided to manually collect a set of audio files. Due to the difficulty to find available actors to record acted sentences, and the more complicated situation to obtain recordings by real-world situations, we collected audio file from movies and TV series, dubbed by Italian professional actors. Differently by others database used in the emotion recognition, in which the number of speakers vary from 5 to 10 like in [10] and [4], our database construction is aimed at creating a generic corpus: 40 movies and TV series are taken into account and, for each of them, sentences acted by different actors are collected. Thus the number of speakers is relatively high, making the system as independent as possible on the speaker. The Italian Emotional Corpus, named *ITA-DB*, is composed by 391 balanced samples of different emotional states that respect Italian judicial proceedings: *anger, fear, joy, sadness and neutral*. This subset of emotions are chosen in order to model the most interesting emotional states, from judicial actors point of view, that could occurs during Italian courtroom debates.

3 Extraction of vocal signatures

Despite there is not yet a general agreement on which are the most representative features, the most widely used are prosodic features, like fundamental frequency (also known as F_0) and formants frequencies (F_1, F_2, F_3), *energy* related features and Mel Frequency Cepstral Coefficients ($MFCC$). Fundamental and formants frequencies refer to the frequency of vocal cords vibration, labelling the human vocal tone in a quite unambiguous way; energy refers to the intensity of vocal signal and Mel Frequency Cepstral Coefficients concern the spectrum of the audio signal. Duration, rate and pause related features are also used, as well as different types of voice quality features. In our work, for each audio file, an attribute extraction process was performed. Initially audio signal was sampled and split in 10ms frames and for each of these frames 8 basic features were extracted. We calculated prosodic features such as F_0, F_1, F_2, F_3, intensity related features like energy and its high and low-passed version and a spectral analysis made up of the first 10 MFCC coefficients normalized by Euclidean Norm. After this first step a 8 features vector for each frame was obtained. In order to extract from this information the necessary features, we considered their respective 3 time series, i.e. the series itself, the series of its maxima and the series of its minimum, and we computed a set of statistical index.

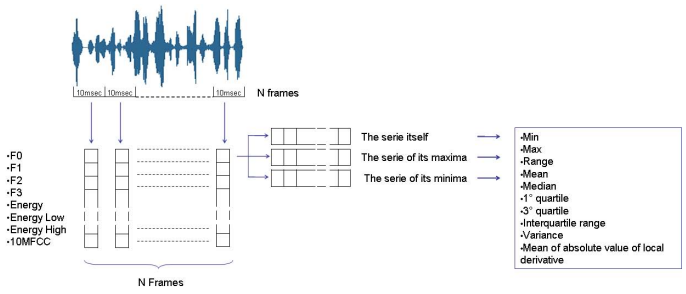


Fig. 2. Feature Extraction Process

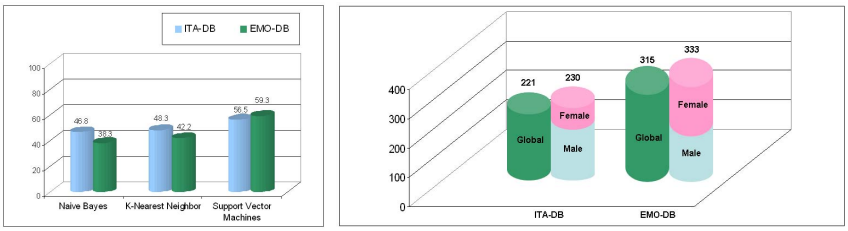
In particular, for each series that describe one of the attribute over the N frames, we computed 10 statistics: minimum, maximum, range (difference between min and max), mean, median, first quartile, third quartile, interquartile range, variance and mean of the absolute value of the local derivative. At the end of this feature extraction process, each vocal signal is represented into a feature space characterized by 240 components ($8 \times 3 \times 10$). In Figure 2 the entire features extraction process is depicted.

4 Emotional State Inference Models

The feature extraction phase, that creates a feature vector for each audio file, allow us to consider emotion recognition as a generic machine learning problem. The learning algorithm investigation, presented in the following subsections, can be distinguished in Flat and Multi-Layer classification.

4.1 Flat Classification

The experimental investigation, show that the machine learning algorithm that performs better is the one based on Support Vector Machines. It is interesting to note that some *similar* emotions (similar in terms of vocal parameters), like anger/joy, neutral/boredom and neutral/sadness, do not allow the classifier to distinguish between them (See Emo-DB in Figure 3(c) and ITA-DB in Figure 3(d)). Another interesting remark, highlighted in Figure 3(b), is related to the investigation about male and female emotion classification performed by two distinct SVMs: learning gender-dependent models produce better performance than unique model. This because some features used to discriminate emotional states are gender-dependent; the fundamental frequency F_0 is one of them: women usually have F_0 values higher than men because of the different size of the vocal tract, in particular the larynx. Starting from this conclusions, we defined a multi-layer model based on the optimal learner, i.e. Support Vector Machines.



(a) (b)

a	b	c	d	e	f	g	← Classified as
102	1	4	6	12	0	2	a = anger (127)
0	56	6	2	0	3	13	b = boredom (80)
6	4	21	7	1	1	6	c = disgust (46)
13	3	7	27	11	1	5	d = fear (67)
31	1	3	10	22	0	4	e = joy (71)
0	6	3	4	0	46	3	f = sadness (62)
1	20	5	6	2	3	41	g = neutral (78)

(c)

a	b	c	d	e	← Classified as
51	5	14	4	6	a = anger (80)
5	44	9	8	5	b = fear (71)
16	10	44	1	9	c = joy (80)
7	7	1	47	18	d = sadness (80)
10	8	6	21	35	e = neutral (80)

(d)

Fig. 3. Flat Classification Experimental Results

4.2 Hierarchical Classification

As highlighted in the previous sections, inference model are influenced by language, gender and "similar" emotional states. For this reasons we propose a Multi-Layer Support Vector Machine approach, that tries to overcome the mentioned limitations. At the first layer a *Gender Recognizer* model is trained to

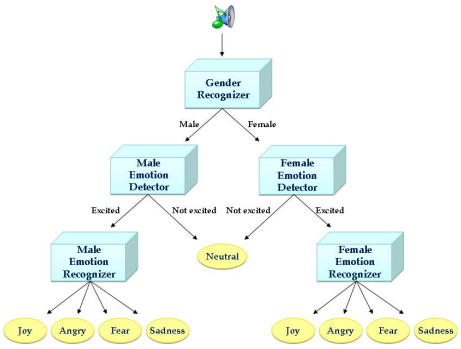
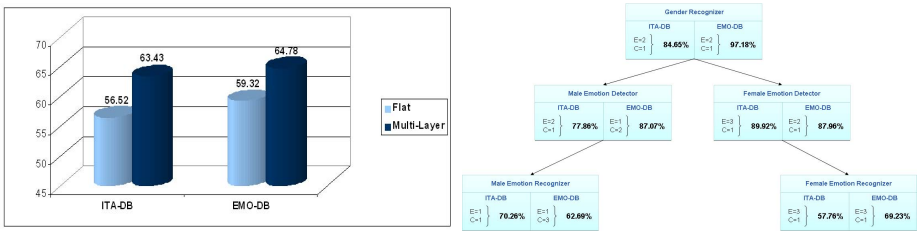


Fig. 4. Multi-Layer Support Vector Machines

determine the gender of the speaker, distinguishing "male" speakers from "female" ones. In order to avoid overlapping with other emotional states, at the second layer gender-dependent models are trained. In particular, *Male Emotion Detector* and *Female Emotion Detector* are induced to produce a binary classifi-

cation that discriminates the "excited" emotional states from the "not excited" ones (i.e. the neutral emotion). The last layer of the hierarchical classification process is aimed at recognizing different emotional state using *Male Emotion Recognizer* and *Female Emotion Recognizer* models, where only "excited" sentences are used to train the models for discriminating the remaining emotional states. A synthetic representation of Multi-Layer Support Vector Machines is depicted in Figure 4. Since also in this case all the models embedded into the hierarchy are based on Support Vector Machines, we experimentally estimate the optimal parameters combination. The performance obtained by the Multi-Layer Support Vector Machines are then compared with the ones provided by the traditional "Flat" Support Vector Machines for both Emo-DB and Ita-DB. The comparison reported in Figure 5(a) highlights the improvement, in terms of number of instances correctly classified, obtained by the Multi-Layer Support Vector Machines with respect to the traditional model. Figure 5(b) shows the



(a) (b)

Fig. 5. Multi-Layer Experimental Results

classification performance of each intermediate layer of the hierarchy. This has been done to understand how the error rate is obtained by the different classifiers of the hierarchy. As we go down in the hierarchy layers the performance get worse, and in the last layer they suffer a remarkable reduction. This because the classifiers have different target: in the root and in the first level, learning is simplified using only two classes, "male" and "female" for root and "excited" and "not excited" for the first layer classifiers; in the last layer a more complex discrimination is required: 6 emotions for Emo-DB and 4 for ITA Emotional DB. A further motivation, related to the decreasing number of instances used to estimate models in the lower layer, could explain the performance reduction. In fact while *Gender Recognizer* can learn on the entire dataset, learning on *Male* and *Female Emotion Detector* is performed on two subsets of the whole dataset, the first model is trained by using only male instances and the second one by considering only female samples. The same thing happens for the last layers, i.e. *Male Emotion Recognizer* and *Female Emotion Recognizer*, that are induced by using "excited" female and "exited" male samples respectively.

5 Conclusion and Future Work

In this paper the problem of producing semantic annotaion for multimedia recording of judicial proceeding is addressed. In particular, two main contributions are given: the construction of an Italian emotional database for Italian proceedings annotation and the investigation of a multi-layer classification system able to recognize emotional states from vocal signal. The proposed model outperforms traditional classification algorithms in terms of instances correctly classified. In our investigation speakers emotion evolution are not considered. We believe that by taking into account the dynamic of emotional process could improve recognition performance. A further development will regard the fusion of different of information sources in order to produce a more accurate prediction.

Acknowledgment

This work has been supported by the European Community FP-7 under the JUMAS Project (ref.: 214306).

References

1. A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth. How to find trouble in communication. *Speech Commun.*, 40(1-2):117–143, 2003.
2. F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss. A database of german emotional speech. In *Interspeech 2005*, pages 1517–1520, 2005.
3. K. Slot J. Cichosz. Application of selected speech-signal characteristics to emotion recognition in polish language. In *Proc. of the 5th International Conf.on signals and electronic systems*, pages 409–412, 2004.
4. V. Petrushin. Emotion recognition in speech signal: Experimental study, development, and application. In *Proc. Sixth International Conf.on Spoken Language Processing (ICSLP 2000)*, pages 222–225, 2000.
5. M. Dordevic M. Rajkovic S. Jovicic, Z.Kasic. Serbian emotional speech database: design, processing and evaluation. In *Proc. of the 9th Conf. on Speech and Computer.*
6. B. Schuller, D. Arsic, F. Wallhoff, and G. Rigoll. Emotion recognition in the noise applying large acoustic feature sets. In *Speech Prosody*, 2006.
7. Björn Schuller, Stephan Reiter, and Gerhard Rigoll. Evolutionary feature generation in speech emotion recognition. In *Proceeding of the 2005 IEEE International Conf.on Multimedia and Expo*, pages 5–8, 2006.
8. M.H. Sedaaghi, C. Kotropoulos, and D. Ververidis. Using adaptive genetic algorithms to improve speech emotion recognition. In *Proc. of 9th Multimedia Signal Processing Workshop*, pages 461–464, 2007.
9. Pierre yves Oudeyer. Novel useful features and algorithms for the recognition of emotions in speech. In *Proc. of the 1st International Conf.on Speech Prosody*, pages 547–550, 2002.
10. Pierre yves Oudeyer. The production and recognition of emotions in speech: features and algorithms. *Int. J. Hum.-Comput. Stud.*, 59(1-2):157–183, 2003.