# Towards a Small Set of Robust Acoustic Features for Emotion Recognition: Challenges

Marie Tahon, Laurence Devillers

# Towards a small set of robust acoustic features for emotion recognition: challenges.

Marie Tahon, Laurence Devillers, *Associate Member, IEEE.*

***Abstract*—The search of a small acoustic feature set for emotion recognition faces three main challenges. Such a feature set must be robust to large diversity of contexts in real-life applications; model parameters must also be optimized for reduced subsets; finally, the result of feature selection must be evaluated in cross-corpus condition. The goal of the present study is to select a consensual set of acoustic features for valence recognition using classification and non-classification based feature ranking and cross-corpus experiments, and to optimize emotional models simultaneously.**

**Five realistic corpora are used in this study: three of them were collected in the framework of the French project on robotics ROMEO, one is a game corpus (JEMO) and one is the well-known AIBO corpus. Combinations of features found with non-classification based methods (Information Gain and Gaussian Mixture Models with Bhattacharyya distance) through multi-corpora experiments are tested under cross-corpus conditions, simultaneously with SVM parameters optimization. Reducing the number of features goes in pair with optimizing model parameters. Experiments carried on randomly selected features from two acoustic feature sets show that a feature space reduction is needed to avoid over-fitting. Since a Grid search tends to find non-standard values with small feature sets, the authors propose a multi-corpus optimization method based on different corpora and acoustic feature subsets which ensures more stability.**

**The results show that acoustic families selected with both feature ranking methods are not relevant in cross-corpus experiments. Promising results have been obtained with a small set of 24 Voiced Cepstral coefficients while this family was ranked in the 2nd and 5th positions with both ranking methods. The proposed optimization method is more robust than the usual Grid search for cross-corpus experiments with small feature sets.**

***Index Terms*—emotion recognition, acoustic features, cross-corpus, SVM parameter optimization, Information Gain, Bhattacharyya distance.**

## I. Introduction

In the field of human-robot interaction (HRI), emotion recognition can be useful for many applications, especially with robotic assistants and social companions. Non-verbal information and particularly audio information play an essential role in human communication [1], [2], but undoubtedly, complementary emotional information exist in other modalities. The authors are involved in the French project on robotics ROMEO[1]. The robot designed in this project faces rich affective interactions in different contexts (speaker's acoustic environments, tasks, real end-users) and can collect inputs from various sensors, such as audio, video, tactile and physiological. According to Batliner et al. [3], obtaining large amounts of real-life data is currently one of the most important hurdles. Since many situations must be represented in real-life data, the collection of real-life emotions requires having large panels of speakers, several different acoustic environments, different emotional and social contexts, which are expensive to set up. Available corpora for emotion recognition in real tasks are nowadays still sparse and contain few instances. Furthermore, most of the available corpora are not yet in the public domain for privacy reasons. Cross-corpus conditions allow to deal with the issue of emotional data availability. Cross-corpus experiments face different challenges: real-life data, rich affective contexts and small corpora.

Because of all these challenges in real-life emotional data cross-corpus experiments, researchers must generalize emotional models as much as possible. The authors' hypothesis is that an appropriate feature space reduction allows generalizing the models, then improves the recognition rate for unseen data [4]. Of course, feature space reduction has also the advantage of being computationally faster, which is highly positive for real-time recognition systems such as robots. Indeed, feature space reduction allows to avoid over-fitting and models tend not to learn by heart the training corpus. Different techniques have already been experimented in affective computing for selecting the best acoustic features. However, few experiments are carried on real-life data in multi-corpus conditions. Automatic feature selection techniques select a subset of features according to a criteria (usually the classification rate itself). This technique has the advantage of being easy to perform in cross-corpus conditions, however it is very time-consuming. Feature ranking techniques are very interesting since they are easy to analyze, but, as of now, such techniques have been used with only one corpus in cross-validation conditions. The originality of the present work lies in the use of two ranking methods to find the most relevant families of features: individual ranking with Information Gain and group of features ranking with Gaussian Mixture Models and the Bhattacharyya distance. Random combinations of features are tested as baselines. The families of features found with these methods are tested under cross-corpus emotion recognition experiments.

In section II, the state of the art on emotional databases, acoustic features and multi-corpus experiments are described. In section III, the authors present the five databases used in their experiments. Section IV summarizes the set of acoustic cues. Section V presents feature space reduction protocols and results. Section VI presents a multi-corpus optimization

M. Tahon and L. Devillers are with the Human-Machine Communication Department, LIMSI-CNRS, 91403 Orsay, France

L. Devillers is with the University Paris-Sorbonne IV, 28 rue Serpente, 75006 Paris, France

[1]http://www.projetromeo.com

of SVM parameters and cross-corpus valence recognition experiments with different subsets of acoustic features. Results and conclusions are drawn in the last section.

## II. State of the art

In this section, the following three topics are addressed: the existing emotional databases (section II-A), the current state of the art on acoustic features for emotion recognition (section II-B) and previous experiments that use more than one database (section II-C).

### A. Emotional databases

This section presents the definition of emotion used in the article, the collection of realistic databases and HRI databases, and finally, segmentation and annotation issues for emotional databases.

*1) Emotion definition:* Emotion is a complex and multidisciplinary field of research and finding a consensual definition is a hard task. According to Scherer [5], the number of human emotions occurring in the context of social interactions is infinite, subtle and often mixed. The complex nature of emotion can be described by discrete labels [6] and continuous dimensions [7], notably valence (positivity), arousal (responsiveness) and control (dominance).

*2) Acted versus realistic databases:* The collection and annotation of emotional databases are crucial issues in emotion recognition and many studies have already been realized in the framework of HUMAINE[2] [8], [9]. A decade ago, emotion recognition models were trained using acted artificial data, such as the Danish Emotional Speech (DES) [10] or the Berlin Emotional Speech-Database (EMO-DB) [11]. The two main drawbacks of the standard corpora used in the community are the very small size of audio corpora and the data variability in terms of task, speaker, age and audio environment which compromise the significance of results and improvements [12]. As a consequence, there is a critical need for data collection with end-users (with different types of speakers, ages) and real tasks for emotion recognition systems since realistic emotions could not be found in acted databases [13]. However, in real-life contexts, emotions are quite sparse, neutral speech is predominant and emotions are shaded [14], [15]. Some prototypical databases are also collected. They contain consensual emotions defined as "utterances that are consistently recognized by a set of human evaluators" [16]. Therefore, their collection with large panels of speakers and different acoustic environments is simplified.

*3) Human-Robot Interaction emotional data:* In order to implement emotional models in systems such as robots, new data must be collected during interactions with these systems. So far, very few HRI databases have been collected. Among them, AIBO database [17] was collected during a child-robot interaction with the AIBO Sony pet. The SEMAINE database [18] was recorded during emotionally rich interactions with an automatic agent. The Herme database [19] was collected in the Science Gallery of Dublin, during conversations between visitors and a robot. Some French emotional HRI databases have also been collected with visually-impaired people (IDV-HR [20]) and children (corpus NAO-HR [21], [22]) interacting with the robot Nao. The collection of such data is more challenging than acted data (protocol, speakers, annotation) and are usually not freely available.

*4) Segmental annotations:* Annotations can be realized at different time level: whole sentence, segment (homogeneous emotion among the segment), word, phoneme, etc. [23]. Therefore the first task is a segmentation task. Segments boundaries can be defined automatically and/or manually. Annotation can be either supervised: categories and/or dimensions are defined before the annotation task, or unsupervised: the annotator chooses the word which best fits the emotion he or she perceives. In the first case, emotional states are constrained and the annotation scheme is highly context-dependent, but finding a consensus between annotators is easy. In both cases, due to the scarcity of data, precise emotion labels have to be mapped onto macro-classes. For example, hot anger is remarkable and is easily annotated as anger, but other anger micro-classes (such as boredom, irritation, ...) are specific to each corpus. Macro-classes are usually driven by emotional theories such as the "big six" theory from Ekman [24]. Valence and Arousal are often annotated on a continuum, but categories are also used [21].

### B. Acoustic features

This section deals with the extraction and normalization of acoustic features. The "holy grail" is to find which features are the best for modeling the emotional speech [25]. The section also presents a review on multi-corpus experiments.

*1) Features reviews:* In the last years, a lot of acoustic features have been developed for speech signal analysis and musical information extraction. Most of them are also used for emotional speech [3], [12], [26]. Acoustic feature sets used in emotion recognition models mainly contain prosodic (pitch, energy, rhythm), timbre and voice quality features. The HUMAINE association also took an inventory of acoustic features in the CEICES initiative [27]. Some features are segmental (they are computed at the frame level), others are supra-segmental (they are computed on a whole sentence, or emotional instance). Some are computed on voiced, others on unvoiced signals [28].

*2) Reduction of the feature space:* The most difficult task in building emotion detection systems is to find the most appropriate acoustic features [29]. The "best feature set" would be small enough to be implemented in a real-time system (not much time-consuming) and robust enough to detect real-life emotions. There are two main approaches for

[2]http://emotion-research.net

this task. The first one consists in extracting all state-of-the-art acoustic features and apply automatic feature selection algorithms to reduce the feature space. Another approach consists in studying and analyzing precisely each feature to find relevant acoustic features. Precise analysis is carried on clean and sparse data and will not be treated in the present paper, whereas feature selection is used on corpora which are collected in different situations. Both approaches are complementary for affective speech study.
Techniques which reduce the feature space dimensionality, consist in a mapping of the input acoustic space onto a less dimensional one, while keeping as much information as possible. The Principal Analysis Component (PCA) is one of the most common technique. PCA reduces the feature space without degrading emotion recognition rate [30]. The interpretation of the selected variables is not easy with PCA since variables are projected variables and do not necessarily have a physical sense.

*3) Feature ranking and selection algorithms:* Feature ranking and selection algorithms have been extensively developed and studied in literature. Filters founded on information theory such as the Information Gain or Best First algorithms have been also used for feature selection even if they perform individual feature ranking [31], [32]. Filters usually have the advantage of being computationally simple. Wrappers selection employ a target classifier's rate as optimization criterion. The Sequential Forward Search technique has been used for emotion recognition [33]. Four feature selection algorithms have been compared by Altun & Gökhan [34], showing that the Least Square Bound Feature Selection algorithm is superior in terms of reducing the average cross-validation error. This study also shows that prosodic and sub-band energy features are the most selected ones by all algorithms. The group-Lasso algorithm does variable selection on (predefined) groups of variables in linear regression models [35]. The advantage of this algorithm is to perform ranking using all features together.

Most of the experiments described in literature are carried on one single corpus. Even if some general trends appear among the selected features, it is quite hard at that time to find a consensus. However, prosodic features – pitch, sound intensity and duration – and cepstral features seem to be highly relevant for emotion recognition [36], [37], [38]. One of the main problem is that the selected features are usually different from one corpus to another and from one task to another and also depend on the selection technique used.

### C. Multi-corpus experiments

*1) Cross-corpus protocols:* Experiments on a single corpus do not enable estimation of the influence of the sources of variability on emotion recognition because they are kept constant. Many studies are carried on one corpus only and researchers cannot generalize the results to other corpora. Cross-corpus experiments consist in using one corpus as training set and using another one as test set. By this way, recognition rates are low but more realistic [39]. Three protocols are predominant in cross-corpus experiments (with $N$ corpora). The leave-one-corpus-out protocol (the "unit" protocol) consists in training on $N-1$ corpus, testing on the last one [40], [41]. The model can be trained with a few corpora and improved with an unsupervised adaptation [42]. It allows merging different data: fixed or variable linguistic information, realistic or acted data, children or adults. In Marchi et al. [43], the authors conclude that accumulation of speech data similar to the testing conditions in the training set improves the recognition performances. The second protocol consists in training on one corpus and testing on the $N-1$ other corpora independently. Then, the recognition rate for each corpus is obtained via the majority voting (the "vote" protocol [44], [**?**]) or average classification rate ([20]) of the $N-1$ classifiers trained independently on the remaining corpora.

Since the Interspeech 2009 Emotion Challenge [45], the evaluation of emotion detection systems has been standardized with the UA (Unweighted Average recall) rate. Schuller et al. [46] performed a cross-corpus test with the "vote" protocol using the acoustic set available for the Interspeech 2012 Challenge using z-normalized (or mean variance normalized) features. For binary valence recognition, the UA results on seven corpora – acted (DES, EMO-DB), induced (eNTERFACE) and realistic (VAM, SAL and SUSAS) – are from $50\%$ to $55\%$. Such an experiment is very interesting to validate a full cross-corpus protocol on merging corpora with a large number of database. Another strategy consists in finding which corpus (or group of corpora) is the best for training the emotional model and which features best fit the final application [36], [47]. The goal of this strategy is to analyze corpora and features, keeping in mind a specific application. In [48], the authors show with cross-corpus experiments (on medical emergency and electricity supply call-centers recordings) that it is possible to generalize a corpus recorded in a certain context to the recognition of emotions elicited in other situations.

Indeed, feature selection results or classification performances obtained with a training set and a test set extracted from the same corpus are usually not robust enough for other corpora. French corpora are rarely used in cross-corpus experiments. Most of the cross-corpus experiments are carried within the same language category because cross-language recognition rates may be even more worse.

*2) Multi-corpus model parameters optimization:* As far as the authors know, there are very few studies on model parameters optimization in emotional cross-corpus experiments. Most studies use Support Vector Models (SVM) and Sequential Minimal Optimization (SMO) function [46], [49]. The kernel parameter $\gamma$ and the complexity $c$ may be optimized using the GRID algorithm or set to fixed values. Optimization is usually performed on the training data, but a development database may be used for optimization purposes [50]. However, Huang and Wang say that "obtaining the

optimal feature subset and SVM parameters must occur simultaneously" [4].

*3) Analyses of acoustic feature distributions:* Acoustic feature distributions across all instances for each affective state of a given corpus, are full of information. In a precedent study, the authors of the present paper introduced a measure of the distance between anger and other emotions for a given acoustic feature [51]. According to this distance, several corpora have been ordered in function of their spontaneity degree. Acoustic features distributions are usually modeled under Gaussian distributions. In Godin & Hansen [52], Gaussian mixtures are used to model physical task stress and neutral speech. In this study, the symmetric Kullback-Leibler distance is used to measure changes in distribution from neutral to physical task stress. The Bhattacharyya distance could measure distance between GMMs to guide phone clustering [53]. Such a protocol is also used for image retrieval [54]. The Bhattacharyya distance is easy to compute for Gaussian distributions and is a powerful tool for dependent ranking features. Such a study is presented later in this paper.

## III. DATABASES

This section presents the five databases used in the following experiments. Four of them are not available for privacy reasons. Therefore, the addition of a publicly available database such as AIBO will enable the scientific community to compare the results. Three of them are collected in the framework of the French project on robotics ROMEO; the prototypical JEMO corpus is collected in the course of an emotion detection game; the AIBO corpus is collected with children interacting with the robot AIBO. This choice guarantees real-life and human-machine interactional data. All lexical contents are free. Two languages are spoken: French and German.

### A. Acquisition protocols of ROMEO databases

The French ROMEO project aims at designing a social humanoid robot. This robot aims to assist elderly and disabled persons at home in everyday life activities, but should also be able to play games with children. The three corpora described below were recorded with a high quality lapel-microphone at the sample frequency of 16 kHz. All participants were native French speakers. The use of these databases carefully respects ethical conventions and agreements, ensuring the anonymity of the speakers, the information privacy and the non-diffusion of corpora and their annotations.

*1) Human-System Interaction with visually impaired people (IDV-HS) [55]:* IDV-HS contains 71 minutes of emotional speech. 28 speakers (11 males and 17 females) were recorded. The recordings took place in a relatively empty room (apart from some basic pieces of furniture) in an apartment block for visually impaired people, who are potential end-users for the ROMEO project. The originality of this corpus lies in the selection of speakers: for a same scientifically controlled recording protocol, we can compare young and elderly voices (from 23 to 89 years old). The experimenter presented six scenarios in which the participant had to pretend to be interacting with a domestic robot called ROMEO. The six scenarios were: medical emergency, suspicious noises, awaking in a good mood, with a bad health, in a bad mood, and finally a visit from close relations. The robot's voice was synthesized on a computer. For each presented scenario, the participant had to picture himself or herself in this context and to speak until the robot detects the right emotion. The emotion detection system was remotely controlled by an experimenter.

*2) Human-Robot Interaction with visually impaired people (IDV-HR) [20]:* The corpus IDV-HR features visually impaired people (young and elderly people) interacting with the humanoid robot NAO from Aldebaran Robotics. 22 speakers from 20 to 80 years old (11 males and 11 females) were recorded in a furnished medical apartment in the IDV, for an amount of 4 hours of recordings. This corpus is collected for the study of both affective states and interactional behaviours. The speaker was asked to play three series of five scenarios (15 scenarios) in which he/she pictured himself/herself in a situation of waking up in the morning. The robot would engage him/her in conversation about either his/her health, or the program of the day, etc. Each of these five scenarios was devoted to a different affective state of the speaker: well-being, minor illness, depressed, medical distress, happiness. Each series of five scenarios differed from the other, by the social attitude of the robot– three attitudes among the followings: friendly, empathetic, encouraging, directive, doubtful or machine-like. The goal of having different social attitudes was to study interactional and social markers. The robot was remotely controlled by an experimenter who selected the social attitudes and the utterances which matched the content of the speaker's speech the best.

*3) Human-Robot Interaction with children (NAO-HR) [21]:* In the NAO-HR corpus, we focused on the role of the robot NAO as a game companion. Ten French children (5 males, 5 females), from 8 to 13 years old, were recorded for a total amount of about 30 minutes of recordings. As a game companion, NAO had to be able to supervise a game, while also being sensitive to the emotions of the children. It should for example be able to detect through the expressed emotions if a child is sad, or if there is a tension between the children. In the NAO-HR corpus, two children by session were recorded as they played with the robot at the LIMSI laboratory. Children were offered to play three games with the robot: a question-answer game, a song game and an emotion-acting game [22]. This corpus has also been collected to study affective and interactional markers during children-robot interactions [56], [57].

### B. Segmentation and annotation scheme of ROMEO databases

Instance's boundaries are defined on each speaker's track. An instance is emotionally homogeneous, i.e. the emotion

Table I
EQUIVALENCES BETWEEN EMOTION LABELS AND VALENCE MACRO-CLASSES

| neutral state | negative state | | | positive state | others |
|---|---|---|---|---|---|
| neutral state | anger<br>irritation<br>scorn | sadness<br>disappointment<br>boredom | fear<br>anxiety<br>stress<br>embarrassment | joy<br>positive<br>amusement<br>satisfaction<br>empathy<br>motherese | irony<br>surprise<br>provocation<br>excitation<br>interest |

is the same and of a constant intensity along the segment [58]. Since instances do not refer to word or sentence level, their durations vary quite a lot as shown in table II. Segmentation is done manually with the Transcriber tool [59] since automatic segmentation using algorithms based on vocal activity detection often makes mistakes [3]. Some parasite noise happened to be audible in the studio (guide dog walking around, people working outside, etc.). When overlapping the speaker's speech, these parts were discarded during the manual segmentation.

Two expert coders perceptively annotated – as far as possible they did not take into account the semantic content – the segment along an annotation scheme inspired by the MECAS (Multi-level Emotion and Context Annotation Scheme) [14]. This scheme enables the representation of complex and realistic emotions using a major and a minor emotion label for each instance. The corpora are annotated with affective state labels which describe each emotional instance (a major and a minor emotional label) and valence and arousal ratings (from $-2$ to 2). Valence dimension describes the sensation expressed by the speaker (positive or negative) and arousal dimension describes the strength of the expressed emotion (weak or intense). The annotation scheme also includes interactional dimensions (for other studies relative to the HRI), the audio quality and whether the participant is acting or talking spontaneously.

In the present paper, only the valence label is used. Valence is known to be one of the most challenging emotional dimension to recognize (in comparison to arousal for example) [46]. It is defined with three different states: neutral, positive and negative (see table I). The negative state contains diverse emotions such as anger and sadness. The "neutral" state contains emotions which have not been labeled as either positive or negative. Only consensual instances are kept for our experiments, therefore the agreement between coders is perfect. The number of emotional segments per class is balanced not to favor a class more than the others. The number of instances used in the present paper and the characteristics of each corpus (total number of consensual instances before balancing, minimum, and mean instance duration, number of speakers and their age, and finally the number of segments after balancing classes) are summarized in table II.

A characteristic of importance is the instance duration after segmentation. Indeed, acoustic features need a minimum number of samples to be computed. Real-life instances' duration lasts less than 1 s whereas acted instances' duration lasts usually more than 1 s, therefore acoustic feature extraction on real-life data usually makes more mistakes. Many studies do not consider instances which last less than 1 s. This is the case in the AIBO corpus, as shown in table II. However these short instances often have strong emotional contents, generally called affect bursts (laughs, cries, etc.) [60].

Table II
CHARACTERISTICS OF THE CORPORA

| Corpus | # total inst. | duration (s) min | duration (s) avg | # speaker (age range) | # 3-class balanced |
|---|---|---|---|---|---|
| NAO-HR [21] | 1275 | 0.23 | 1.46 | 12 (8-13) | 675 |
| IDV-HS [55] | 2898 | 0.18 | 1.73 | 28 (23-89) | 1656 |
| IDV-HR [61] | 6072 | 0.24 | 2.45 | 22 (20-80) | 2394 |
| JEMO [62] | 1937 | 0.27 | 1.83 | 64 (21-45) | 1485 |
| AIBO-O [17] | 2252 | 1.00 | 2.34 | 26 (10-13) | 1497 |
| AIBO-M [17] | 1738 | 1.00 | 2.21 | 25 (10-13) | 1107 |

### C. Benchmarks corpora

*1) JEMO corpus [62]:* The JEMO corpus was collected in the course of the ANR (French National Research Agency) Affective Avatar project. It was recorded in order to obtain prototypical emotions. The speakers were asked to play an emotion-detection game during which the system had to recognize their emotion. The lexical content was totally free and the language is French. The system used a speech segmentation based on silences and pauses and an emotion recognizer based on four emotions: neutral, anger, sadness and positive (or joy). The system detected the emotions from the audio signal and presented the detected emotion on a screen visible to the player. 64 speakers were recorded (29 men, and 35 women) for a total duration of about 30 minutes. This corpus contains mainly prototypical emotions, i.e. emotions clearly recognizable by coders. In the present paper, the four emotional macro-classes have been grouped into three valence states according to table I. The number of instances used in the following experiments and the characteristics of the corpus are presented in table II.

*2) AIBO corpus [17]:* The FAU AIBO corpus was collected during an interaction between a child (from 10 to 13 years old) and the Sony dog robot called AIBO. The German lexicon of the AIBO corpus comprises words that occur very

frequently. The recordings took place in two different schools: in a class room (Mont) and in a playground (Ohm). In order to have similar acoustic environments in each corpus, the recordings of the Ohm school (AIBO-O) and the Mont school (AIBO-M) are split in two. The authors use two different configurations of the AIBO corpus. The CEICES configuration [27] is used for multi-corpus experiments (family selection, model parameters optimization and cross-corpus experiments). In this configuration, emotional instances are annotated with anger, motherese, empathy and neutral labels. The balanced number of emotional segments per class used in the present paper, is reported in table II. The authors also use the Interspeech 2009 configuration [45] as a baseline to compare the reference acoustic feature set OpenSmile 2009 and the acoustic feature set used in this paper (see section IV-C). In the 2-class Interspeech challenge, only anger and idle (other emotional states) labels are used. In this configuration, classes are unbalanced. If then, the Weighed Average recall (WA) may be biased. For example all instances are recognized as neutral, the biggest class. This point is very important to be as close as possible to final applications.

## IV. Acoustic cues

The authors developed their own feature set (Li-174), all experiments on acoustic family selection are carried on this specific set. However, for comparison purposes, they also use the baseline OpenSmile set (Os-384). Both feature sets are described in this section.

### A. The 174 acoustic features set

The robot ROMEO will be endowed with a segmentation and feature extraction tool able to detect voice activity, and to compute acoustic features. Since the final goal is to embed the robot with an emotion recognition tool, this tool should extract the minimum number of acoustic features. Each segment is segmented on 30 ms frames every 10 ms using a Hamming window. Low-Level Descriptors (LLD) are extracted from the speech signal at the frame-level. Voiced and unvoiced features have been shown to be relevant for fear-type emotion recognition [63], LLD are extracted on voiced, unvoiced and the full instance. Statistical functions, or functionals, are applied to LLD, thus giving acoustic features at the segment-level. All acoustic cues are normalized to speaker with the Mean-Variance normalization.

174 acoustic features are referenced in the Li-174 set (table III). The choice of the features is inspired by existing feature sets (OpenSmile[3], Yaafe[4], etc.) but also from musical information extraction (perceptive features) [64]. They are grouped in seven acoustic families according to the LLD from which they are extracted. For example the Pitch family contains all functionals computed on the pitch: mean, standard deviation (std), maximum and minimum.

[3]http://opensmile.sourceforge.net/

[4]http://yaafe.sourceforge.net/

### B. Detailed description of the acoustic set

*1) Pitch ($F_0$):* is extracted with Praat[5] [65] each 10 ms. The fundamental frequency is converted in semitones (st) since this scale is close to the human ear perception of frequencies. Praat tool also gives the pulses. Pulses define voiced and unvoiced parts of the speech signal. To limit the number of pitch errors, voiced parts which last less than 40 ms are discarded. For each voiced part, the Li-174 computes mean, std, minimum and maximum statistics on pitch. The mean values of these four statistics are given for each emotional instance.

*2) Energy:* has different definitions. To be as close as possible to the perceptive energy, the global loudness [64] is extracted in Li-174 with Matlab. Global loudness corresponds to the energy of the signal convolved with a perceptive filter (here a Bark filter). For voiced, unvoiced parts and the full signal, the Li-174 computes mean, std, minimum and maximum statistics on loudness.

*3) Spectral features (SPE):* usually describe the spectral envelope. Many methods exist for extracting the spectral envelope (true-envelope [66], Hilbert-Huang transforms [67], etc.). The authors of the present paper chose to compute the LPC envelope, since it has shown its interest for emotion recognition [68]. Roll-off frequencies, spectral Centroid and spectral Slope which describe the shape of the envelop [64], are defined as follows ($a(f)$ is the amplitude of the signal at the frequency $f$):

- Roll-off frequency $f_{ro}$ at $x\%$: $\sum_0^{f_{RO}} a^2(f) = x \sum_0^{f_e/2} a^2(f)$.
- Spectral centroid $f_c$: $f_c \sum_0^{f_e/2} a(f) = \sum_0^{f_e/2} f \cdot a(f)$.
- Spectral slope $s$ is defined according to the linear regression of the spectrum: $a(f) = s \cdot f + c^{te}$ for $f \in [0; f_e/2]$.

*4) Energy bands features (EBB):* are recently used in emotion recognition. Harmonic bands correspond to the spectral energy in the band $\left[\frac{n \cdot F_0}{2}; \frac{(n+1) \cdot F_0}{2}\right]$ with $n$ from 1 to 5 [69]. Bark bands correspond to the spectral energy in the first 21 Bark bands [64]. Bark scale is interesting because it relies on human perception. Harmonic bands are interesting because they focus on the spectral part of most energy.

*5) Mel Cepstral coefficients or MFCC (CEP):* and derivative cepstral coefficients are usual features for emotion recognition. The Mel-cepstrum is the Discrete Fourier Transform of the logarithm of the Melbands spectrum. The authors use the following Mel scale: $M(f) = 1000\left(1 + \log_{10}\frac{f}{1000}\right)$ for $f > 1000$ Hz, $M(f) = f$ for $f < 1000$ Hz [70]. They are known to be robust to noisy signal. Many recent studies show their interest

[5]http://www.praat.org

Table III
THE LI-174 ACOUSTIC FEATURES SET

| Family | LLD | Funcationals | Voiced | Unvoiced | All |
|---|---|---|---|---|---|
| **Pitch (4)** | | | **4** | | |
| | $F_0$ (st) | mean/std/max/min | 4 | | |
| **Energy (12)** | | | **4** | **4** | **4** |
| | Total loudness (Bark, dB) | mean/std/max/min | 4 | 4 | 4 |
| **Spectra (14)** | | | **7** | **7** | **0** |
| | Roll Off 5, 25, 50, 75, 95% (Hz) | mean | 5 | 5 | |
| | Total slope | mean | 1 | 1 | |
| | Centroid | mean | 1 | 1 | |
| **Band energy (47)** | | | **26** | **21** | **0** |
| | Bark bands 0-21 | mean | 21 | 21 | |
| | Harmonic bands 0-5 | mean | 5 | | |
| **Cepstre (78)** | | | **26** | **26** | **26** |
| | MFCC 0-12 | mean | 13 | 13 | 13 |
| | $\Delta$MFCC 0-12 | mean | 13 | 13 | 13 |
| **Formants (14)** | | | **14** | **0** | **0** |
| | $F_1$ (st) | mean/std/max/min | 4 | | |
| | $F_2$ (st) | mean/std/max/min | 4 | | |
| | $F_3$ (st) | mean/std/max/min | 4 | | |
| | Articulation dist. | mean/std | 2 | | |
| **Voice Quality (5)** | | | **0** | **0** | **5** |
| | PunvoicedPraat | | | | 1 |
| | JitterLocalPraat | | | | 1 |
| | ShimmerLocalPraat | | | | 1 |
| | HNRPraat | | | | 1 |
| | Harmonics number | | | | 1 |
| **Total (174)** | | | **81** | **58** | **35** |

for emotion recognition [36], [16], [71].

*6) Formants:* are usual features in emotion recognition. They have been implemented in the 2010 Interspeech Challenge [72]. They are also present in many studies on emotions [14], [73]. Formant-based features such as articulation distance (which consists in an approximation of the area of the vocalic triangle), have been studied for emotion recognition in [47], [74], [75].

*7) Voice quality (VQ):* consists in many high-level features which were developed for voice transformation, speech synthesis, clinical purposes and emotion recognition. Among those descriptors, Harmonic to Noise Ratio (HNR), jitter, shimmer and the voiced vs. unvoiced ratio (punvoiced) have been shown to be relevant for valence recognition [61]. In the Li-174 feature set, HNR, jitter, shimmer and punvoiced are extracted with Praat. The harmonic number corresponds to the ratio: $\frac{f_{ro}(95\%)}{f_{ro}(5\%)}$. This feature estimates the number of harmonics in voiced parts of speech energy.

### C. Baseline

Since the Interspeech 2009 Emotion Challenge, the 384 OpenSmile acoustic set (Os-384) is considered as a reference. It consists in 16 LLD (fundamental frequency, RMS energy, Zero-Crossing Rate, Harmonics to Noise Ratio and 12 cepstral coefficients), their derivatives and 12 functionals (mean, variance, kurtosis, skewness, extremes functions and slope regression coefficients).

A comparison between the reference set Os-384 and the authors' set Li-174 is realized based on the challenge IS09 configuration: two non-balanced classes NEG (2465 instances) and IDL (5792 instances) with WEKA[6]. The training is performed with Ohm school instances and testing with Mont school instances. Results are shown employing the whole feature sets Os-384 and Li-174 with support vector classification (Sequential Minimal Optimization learning, linear kernel, normalization of training data): UAR (Os-384) = 63.8%, WAR (Os-384) = 73.0%, UAR (Li-174) = 60.5%, WAR (Li-174) = 71.7%. Results obtained with Os-384 are slightly better than with Li-174. One of the reason of the difference may be the number of features in each set. Because the number of features in Li-174 is twice less than in Os-384, the classification rate obtained with Li-174 is considered as relevant and will be used in the following experiments.

## V. MULTI-CORPUS FEATURE SPACE REDUCTION

In order to reduce the feature space dimensionality, the authors choose to run two non-classification based ranking protocols and one random ranking protocol. The Information Gain (IG) protocol treat each feature independently while a combination of Gaussian Mixture Models and Bhattacharyya distance (G+B) treat family features as dependent groups. Each ranking experiment are applied independently on each of the six sub-corpora (cross-corpus) and on all merged corpora (merged).

### A. Ranking feature families with Information Gain (IG)

*1) Protocol:* Each of the 174 attributes are continuous values. The WEKA's Information Gain Attribute Evaluation

[6]http://www.cs.waikato.ac.nz/ml/weka

function uses the Minimum Description Length Principle Criterion [76] to find the best cut-points. The Information Gain is the reduction in entropy obtained with the optimized cut-points. Best attributes lead to little cut-points and a small entropy; their information gain is maximum: $IG = 1$ [77].

The feature ranking is ran on the six sub-corpora at our disposal. This algorithm is applied to a three states valence and gives one value per segment and per attribute. For each corpus, the information gain $IG(a)$ is attributed to each feature $a$.

The $IG(f)$ of an acoustic family is obtained while summing all $IG(a)$ of the features belonging to the family (eq. 1). The relative information gain $IG^r(f)$ is normalized to the sum of the 174 $IG(a)$. The main drawback of this presentation is that the number of features per family differs, but what is time-consuming is the extraction of the LLD, not the computation of functionals. Therefore the number of features belonging to the same family is not as important as the LLD. That is why the authors chose to compute the sum and not the mean information gain of all features belonging to the same family. Therefore, this protocol allows to identify best LLDs, not best features.

$$IG(f) = \sum_{a=1}^{N(f)} IG(a) \tag{1}$$

*2) Results:* Multi-corpus (corpus-independent and merged corpora) results are reported in table IV. Differences in ranking between Energy Band (Bark and Harmonic bands) and Cepstral families is very small for IDV-HS, JEMO and AIBO-O. For the other sub-corpora, Energy band family clearly reaches the first place. Unexpectedly, Voice Quality family has a small $IG^r$, while it was shown to be useful for valence recognition [61]. Formants and Pitch have also a small $IG^r$. Pitch family is slightly more important in JEMO ($IG^r = 9.5\%$) and NAO-HR ($IG^r = 7.1\%$) compared to other corpora, probably because in both corpora, speakers were playing games which elicited more excitation. The trends observed on merged corpora results are similar to the ones observed on each corpus independently. More experiments show that Unvoiced and All features are often badly ranked for all corpora.

### *B. Ranking feature families with GMM and Bhattacharyya distance (G+B)*

*1) Protocol:* Assuming that acoustic feature distributions are under Gaussian distributions, each valence class may be modeled using Gaussian mixtures (GMM) with random initialization. Then, the variance of the Gaussian model is a relevant factor to estimate if emotional classes are acoustically homogeneous or not. Experiments with 1 to 256 Gaussians show that 8 Gaussians seem to be enough for the present study. The Bhattacharyya distance [78] is a theoretical distance measure between two Gaussian distributions. The Bhattacharyya distance between two Gaussians $d_{B,i,j}(f)$ for a set of features $f$ (feature family), is given by equation 2, where $M_i$, $M_j$ (resp. $\Sigma_i$, $\Sigma_j$) are the mean matrices (resp. variances matrices) of the two Gaussian distributions $i$ and $j$. The first term is linked to the well-known Mahanalobis distance which is null if the means are similar, even if the variances are different. On the contrary, the Bhattacharyya distance is not null when variances are different. It has the properties of being computationally simple, and to be extensible to Gaussian mixtures. The distance between classes $C1$ and $C2$ is the sum of the different Gaussians $i$ and $j$ according to their weights $W_i$ (equation 3) [54].

$$d_{B,i,j}(f) = \frac{1}{8}\left(M_j - M_i\right)^T \left[\frac{\Sigma_i + \Sigma_j}{2}\right]^{-1} \left(M_j - M_i\right) + \frac{1}{2}\ln\left[\frac{\det\left(\frac{\Sigma_i + \Sigma_j}{2}\right)}{\sqrt{\det(\Sigma_i)\cdot\det(\Sigma_j)}}\right] \tag{2}$$

$$d_{B,12,GMM}(f) = \sum_{i=1}^{N^1_{GMM}} \sum_{j=1}^{N^2_{GMM}} W_i \cdot W_j d_{B,i,j}(f) \tag{3}$$

The final distance $d_B(f)$ is the logarithm of the mean of the three distances Neu/Neg, Neu/Pos and Neg/Pos. The relative Bhattacharyya distance $d^r_B(f)$ is the Bhattacharyya distance $d_B(f)$ normalized by the sum of all family distances. Because the GMM initialization is performed randomly, all results are averaged over 5 folds. For both ranking method, there is an impact of the feature family $f$ dimension on the the family rank.

*2) Results:* The obtained results using the Bhattacharyya distance are presented in table IV. Acoustic family that reaches the highest relative distance is the Spectral family for all corpora independently and merged corpora. Loudness, Cepstral and EBB families just follow the Spectral family. It appears that Pitch and Voice Quality have not a high mean distance, it probably means that emotional classes are not acoustically homogeneous among these families. Formant family is relatively badly ranked when corpora are taken separately, whereas it reaches the fourth position, just before the Cepstral family (fifth position) when corpora are merged. Further results show that Unvoiced features are very interesting to discriminate valence classes in the two AIBO sub-corpora. Unvoiced features are slightly better ranked than Voiced features for all the four French corpora and all merged corpora.

Some general trends exist among the six corpora, but each corpus has its own acoustic specificity. This confirms the fact that acoustic patterns of each valence class are different from one corpus to another, which was expected because of the differences in the recording and annotation protocols.

Table IV
FAMILY RANKING OVER THE SIX SUB-CORPORA AND ALL MERGED CORPORA USING IG AND G+B WITH LI-174 SET.

| Corpus (#features) | NAO-HR | IDV-HS | IDV-HR | JEMO | AIBO-O | AIBO-M | Merged |
|---|---|---|---|---|---|---|---|
| | Relative Information Gain $IG^r$ | | | | | | |
| $F_0$ (4) | 7.1 | 2.2 | 4.4 | 9.5 | 2.5 | 2.5 | 5.2 |
| Loudness (12) | 14.2 | 10.5 | 13.1 | 10.8 | 12.3 | 11.1 | 12.5 |
| Spectral (14) | 5.5 | 10.9 | 7.5 | 12.5 | 11.2 | 10.7 | 12.4 |
| Cepstral (78) | 16.4 | 31.2 | 18.3 | **31.1** | 32.9 | 29.9 | 24.7 |
| EBB (47) | **55.1** | **37.9** | **52.0** | 30.1 | **34.7** | **39.9** | **39.1** |
| Formant (14) | $< 0.001$ | 4.1 | 1.4 | 2.3 | 2.1 | 2.8 | 2.0 |
| VQ (5) | 1.8 | 3.1 | 3.3 | 3.8 | 4.4 | 3.2 | 4.0 |
| Voiced (81) | **63.1** | **46.0** | **72.8** | **57.8** | **69.1** | **74.5** | **70.3** |
| Unvoiced (58) | 20.3 | 33.6 | 13.2 | 15.9 | 17.1 | 12.3 | 9.9 |
| All (35) | 16.6 | 20.5 | 14.0 | 26.3 | 13.8 | 13.3 | 19.8 |
| | Relative Bhattacharyya distance $d^r_{B,GMM}$ | | | | | | |
| $F_0$ (4) | 6.8 | 4.1 | 6.1 | 6.7 | -2.8 | -2.8 | 8.3 |
| Loudness (12) | 16.9 | 17.5 | 17.3 | 15.7 | 22.5 | 20.4 | 16.1 |
| Spectral (14) | **19.7** | **23.4** | **21.0** | **23.4** | **31.6** | **34.7** | **26.7** |
| Cepstral (78) | 19.3 | 19.7 | 19.7 | 18.9 | 26.2 | 25.4 | 14.5 |
| EBB (47) | 19.4 | 20.1 | 19.6 | 18.7 | 26.0 | 25.0 | 16.8 |
| Formant (14) | 12.5 | 11.6 | 10.9 | 10.7 | -0.9 | -0.8 | 14.9 |
| VQ (5) | 5.5 | 3.7 | 5.5 | 6.0 | -2.6 | -2.0 | 2.7 |
| Voiced (81) | 40.5 | 43.9 | 43.1 | 38.2 | 35.9 | 17.6 | 40.6 |
| Unvoiced (58) | **49.1** | **45.0** | **47.9** | **40.6** | **59.7** | **76.1** | **42.0** |
| All (35) | 10.4 | 11.0 | 9.0 | 21.2 | 4.3 | 6.2 | 17.5 |

Table V
DESCRIPTION OF THE FEATURE SUBSETS

| Name | Origin | Description |
|---|---|---|
| Os-384 | OpenSmile | Full OpenSmile set |
| Os-R50 | | Mean random combination of 50 features (3 sets) |
| Os-R25 | | Mean random combination of 25 features (3 sets) |
| Os-CEP24 | | 24 OpenSmile Cepstral amean coefficients (1-12 and derivatives) on voiced |
| Li-174 | Limsi | Full Limsi set |
| Li-R50 | | Mean random combination of 50 features (3 sets) |
| Li-R25 | | Mean random combination of 25 features (3 sets) |
| Li-B50 | | Best 50 features obtained on the training corpus with InfoGain individual ranking (6 sets) |
| Li-B25 | | Best 25 features obtained on the training corpus with InfoGain individual ranking (6 sets) |
| Li-EBB47 | | 47 Energy Bands (Bark and Harmonic bands) on voiced and unvoiced |
| Li-CEP48 | | 48 Cepstral coefficients (1-12 and derivatives) on voiced and unvoiced |
| Li-EBB26 | | 26 Energy bands on voiced |
| Li-CEP24 | | 24 Cepstral coefficients (1-12 and derivatives) on voiced |
| Li-SPE14 | | 14 Spectral features on voiced and unvoiced |

### C. Subset of features

Since they aim for to feature space dimensionality reduction, the authors chose to select the 3 first-ranked families extracted on Voiced and Unvoiced signals, which are similar with both IG and G+B ranking algorithms: Cepstral (48 features, coefficient 0 is removed for comparison purposes with OpenSmile feature set), Spectral (14 features) and Energy bands (47 features). Since the number of feature per family highly differs, the dimension is reduced again while selecting only voiced features. In order to compare a selection based on acoustic families of features and independent features, 50 [resp. 25] best-ranked features with IG, are selected for each corpus, forming the Li-B50 [resp. Li-B25] feature set. For testing the hypothesis "the system is over-fitting when using too much features", random features sets have been tested on both Li-174 and Os-384 full sets. The authors chose to add random feature sets to have a baseline study [79]. The number of feature in random sets is defined by the number of features in selected sets. The different subsets of features are reported in table V. They consist in full sets, selected family and best independent features sets, random sets, selected from Li-174 and Os-384 sets. The robustness of these subsets will be validated in the next section with cross-corpus experiments.

## VI. CROSS-CORPUS EXPERIMENTS

In the following experiments, the cross-corpus emotion classification is performed on the combination of features described in the previous section. Two multi-corpus optimization methods are tested on the six sub-corpora with two acoustic feature sets (Li-174 and Os-384) and randomly reduced feature subsets. Auto-coherence and cross-corpus results are presented, underlying the relevancy of Li-CEP24 subset.

### A. Cross-corpus protocol

Classification is realized with Support Vector Machine (SVM) and Sequential Minimal Optimization function (SMO) since this configuration has been widely used. The authors use LibSVM tool[7]. To better understand the obtained results and

[7]https://www.csie.ntu.edu.tw/~cjlin/libsvm/

to estimate the generalization power of each of the six sub-corpora, the authors add a confidence interval $CI$. It is defined by the relation 4, where $p$ is the prediction score and $N$ the total number of instances [80].

$$P = p \pm CI = p \pm 1.96\sqrt{\frac{p(1-p)}{N}} \quad (4)$$

In cross-corpus experiments, one corpus is used for training and the others for testing. For each experiment, $4 \times 4$ classification rates are obtained. The cross-corpus result consists in the mean value of the classification rates for test with different corpora for training and testing ($4 \times 3$ tests). The auto-coherence experiments (training and testing on the same corpus - 6 tests) are treated separately. Results are given in terms of mean Unweighted Average Recall (UAR %). Since classification experiments are run on three classes, the chance-level is 33.3%.

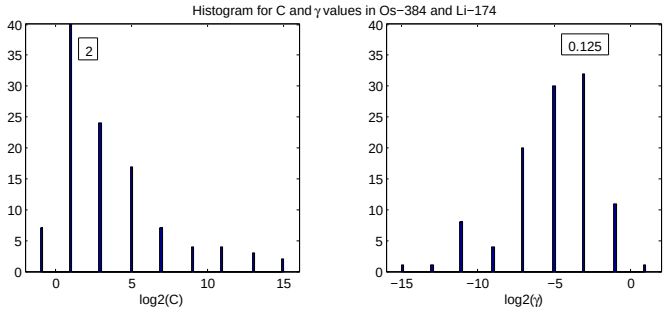
### B. Multi-corpus optimization of SVM parameters

As Huang and Wang pointed out [4], "obtaining the optimal feature subset and SVM parameters must occur simultaneously", this section is thus dedicated to optimization of cross-corpus valence recognition. Obtaining the optimal and minimal feature subset and optimizing of SVM parameters for modeling emotion classes are jointly studied.

*1) Optimization with the Grid search algorithm (OPT1):* The Grid search algorithm is a method for finding the optimal $C$ and $\gamma$ when using the RBF kernel function. Although this method is very well-known, it also has disadvantages: it is time-consuming and do not perform well in all situations [4]. The Grid search selects the couple of parameters $(C;\gamma)$ that leads to a maximum of accuracy (accuracy is averaged over n-folds classification tests) for a given training database. However, other local maxima could be eligible if a confidence interval was given. Non-standard values of $(C;\gamma)$ tend to be reached by the Grid search algorithm when the number of feature is too small (typically less than 50) with both Li-174 and Os-384 full sets. Selected $(C;\gamma)$ are often non-standard values with ramdomly reduced feature subsets, but it is not the case with appropriate feature family. Such non-standard values usually lead to a serious drop in cross-corpus classification performance.

*2) Fixed optimized values (OPT2):* In order to ensure SVM parameter stability and to avoid non-standard values of $(C;\gamma)$, other optimization techniques must be tested. One could use a development database and optimize $(C;\gamma)$ on the accuracy obtained while testing on this development database. The obtained results with the corpora described in section III are not successful, probably because development data is not close enough to the final tested data. Moreover, using such a technique in real-life applications implies to optimize and train again the model on line at each changing situation. Eyben et al [25] propose to average the accuracy over different values of complexity $C$ for the results to be more stable. The authors discarded non standard (extreme) values since "performance drops significantly with the smaller feature set". This confirms the fact that optimization with small feature sets usually lowers classification because the selected $(C;\gamma)$ parameters are often non-standard values. This technique is not adapted in real-life applications either.

Figure 1. Histograms of $C$ and $\gamma$ values obtained with Grid search on Os-384 (weight:3), Os-R50 (3 random sets), Os-R25 (3 random sets) and Li-174 (weight: 3), Li-R50 (3 random sets), Li-25 (3 random sets) over the six sub-corpora.




The optimization technique (OPT2) proposed in this paper should ensure more stability of the parameters across corpora and avoid non-standard values, thus leading to good performances in cross-corpus. $(C;\gamma)$ couple is chosen among the optimized values found by the Grid search over a large number of experiments including different feature sets and corpora. Figure 1 shows histograms of optimized values for both Li-174 and Os-384 and 50, 25 randomly selected features - $7 \times 2$ subsets. The most frequent values are $(C;\gamma) = (2;0.125)$. Such a technique has the advantage of being easy to use in real-life applications.

Table VI
COMPARISON BETWEEN TWO SVM OPTIMIZATION TECHNIQUES. CROSS-VALIDATION (CV) AND CROSS-CORPUS (XC) ACCURACY IN % (WITHOUT AIBO). $CI = \pm 2.5\%$

| Set | OPT1 | | | | OPT2 | | | |
|---|---|---|---|---|---|---|---|---|
| | CV | XC | | | CV | XC | | |
| | avg | avg | min | max | avg | avg | min | max |
| Os-384 | 57.0 | 43.7 | 33.8 | 52.5 | 57.1 | 40.0 | 33.3 | 52.5 |
| Os-R50 | 52.7 | 43.1 | 36.3 | 48.6 | 52.2 | 43.6 | 37.9 | 48.6 |
| Os-R25 | 51.6 | 42.7 | 37.0 | 47.9 | 50.9 | 42.6 | 37.3 | 48.2 |
| Li-174 | 58.4 | 41.2 | 33.3 | 49.6 | 57.7 | 40.0 | 33.3 | 47.6 |
| Li-R50 | 54.2 | 39.9 | 32.9 | 49.5 | 53.1 | 40.4 | 33.4 | 48.3 |
| Li-R25 | 52.8 | 39.0 | 34.2 | 47.9 | 51.3 | 40.9 | 34.4 | 50.5 |

*3) Comparison between the SVM optimization techniques:* Cross-corpus experiments have been realized with the six sub-corpora and two acoustic feature sets. The authors choose to use both their own feature set Li-174 and Opensmile feature set Os-384 for generalization purposes. The average accuracy obtained in cross-corpus conditions with all six sub-corpora is 39.4% with Os-384 and *OPT1*, while it reaches 43.7% without the two AIBO sub-corpora. Since the cross-corpus performance significantly degrades when using AIBO sub-corpora with all subsets - different idioms, mismatches

between emotion classes - the results are presented with the four French corpora only.

Table VI summarizes results obtained in cross-validation (10 folds, training and testing are realized with the same corpus) and cross-corpus conditions (training and testing are realized with different databases) with different subsets of features and the two SVM parameters optimization techniques described previously. For each subset of feature and optimization technique, $4 \times 3$ cross-corpus tests and $4$ cross-validation tests are run. Average, maximum and minimum values are reported in table VI. All the trends described in this section are similar with AIBO, but performances are lower.

Figure 2. Auto-coherence results on different acoustic sets without the two AIBO sub-corpora. $CI = \pm 2.3\%$. Results are given in terms of mean UA (%) over the corpora.

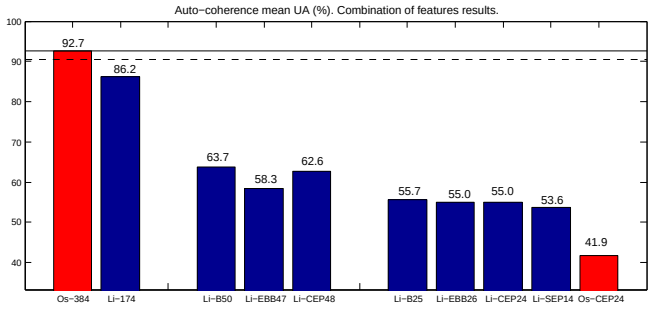
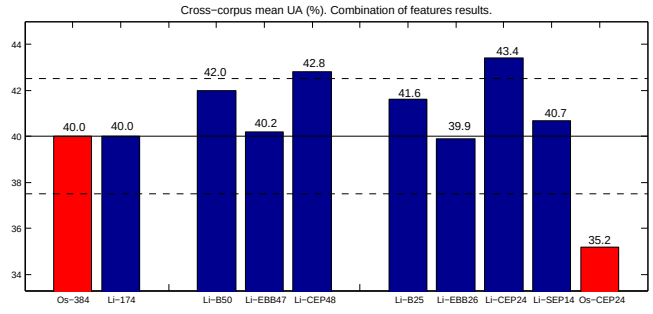



Figure 3. Cross-Corpus results on different acoustic sets without the two AIBO sub-corpora. $CI = \pm 2.5\%$. Results are given in terms of mean UA (%) over the corpora.




Cross-validation (CV) performances are slightly better with *OPT1* optimization than with *OPT2* (differences $< 2.5\%$). For instance with Os-384 [resp. Li-174], *OPT1* gives a CV rate of 57.0% [resp. 58.4%], while *OPT2* gives a CV rate of 57.1% [resp. 57.7%]. Cross-corpus (XC) performances are better on full sets with *OPT1* - Os-384 gives 43.7% [resp. Li-174 gives 41.2%] - than with *OPT2* - Os-384 gives 40.0% [resp. Li-174 gives 40.0%]. XC accuracy degrades with *OPT1* while reducing the number of features, whereas it upgrades with OPT2 with both feature sets. *OPT1* does not perform as well as *OPT2* when the set of features is reduced especially with the Limsi set, for the reasons already mentioned in section VI-B1.

With Os-384 [resp. Li-174], cross-corpus performance with 25 randomly selected features and *OPT2* is 42.6% [resp. 40.9%] are almost similar to the result obtained with full set and *OPT1*: 43.7% [resp. 41.2%]. With an appropriate optimization technique (here the *OPT2*), a random selection of 50 and 25 features performs almost as well as the full sets. The fact that randomly selected feature sets have similar performance than full feature set gives a piece of evidence that the system is over-fitting when using too many features. Therefore, an intelligent feature selection should significantly outperform performances obtained with full sets.

### C. Auto-coherence results

In this section, the authors use the same corpus for training and testing under auto-coherence experiments. Such an experiment gives a baseline of the most important rate a system could achieve and how this system recognizes the data it has already seen [81]. Figure 2 shows that the highest the number of features, the better the recognition rates. With few features the auto-coherence is closer to cross-corpus results. In this specific case, optimization of SVM parameters significantly improves the recognition rates. The IG individually best ranked features Li-B50 [resp. Li-B25] gives the best recognition rate at 63.6% [resp. 55.7%] among the reduced feature subsets with almost 50 features [resp. 25 features]. One of the reasons is that the IG individually best features were ranked on each training corpus.

### D. Cross-corpus results

Since the authors' goal is to find a small acoustic feature set, the following experiments are realized in cross-corpus conditions, SVM parameters are set to the fixed optimal values obtained with the *OPT2* method. Cross-corpus results are reported in figure 3. IG individually best ranked subset of features Li-B50 [resp. Li-B25] gives good recognition rate at 42.0% [resp. 41.6%] among the reduced feature subsets with almost 50 features [resp. 25 features], however improvements from random subsets and full sets are not significant ($< 2.5\%$). IG and G+B best ranked families (Li-EBB47 best ranked with IG and Li-Spec14 best ranked with G+B) do not give interesting improvements from full sets. However such reduced feature sets do not degrade recognition rates. It means that ranking such as Information Gain and Gaussian Models with Bhattacharyya distance is not powerful enough for selecting features in cross-corpus experiments.

Cepstral family which was ranked in second position with IG and fifth position with G+B, gives the best recognition rates (Li-CEP48: 42.8% and Li-CEP24: 43.4%) and significant improvements from full feature sets. The result obtained with the Li-CEP24 set (43.4%) is significantly higher than the results obtained with the full sets Li-174 (40.0%) and Os-384 (40.0%), and with the cepstral reduced Os-CEP24 set (35.2%). In both Li-174 and Os-384, cepstral features are extracted with the same algorithm but the Mel scales slightly differ:

OpenSmile (eq. 5) is based on HTK toolkit [82] and Li-174 (eq. 6) is based on Rabiner and Juang [70].

$$M(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{5}$$

$$M(f) = \begin{cases} 1000\left(1 + \log_{10}\frac{f}{1000}\right) & \text{for } f > 1000\text{Hz}, \\ f & \text{for } f < 1000\text{Hz} \end{cases} \tag{6}$$

The authors conclude that the implementation used in the Li-174 for extracting cepstral coefficient gives promising results for cross-corpus valence recognition.

With a large number of features, models fit perfectly with the training data; auto-coherence rates are very high and cross-corpus rates are quite low. With a smaller set of features, models tend to generalize; auto-coherence rates degrade while cross-corpus rates may be improved. This result confirms Huang and Wang [4] and heighten the relevancy of small feature sets for automatic real-life valence recognition. Of course, this result is pertinent once model parameters are correctly optimized. Similar trends occur when adding the two AIBO sub-corpora, the only difference is that performance rates are lower.

## VII. Conclusion

The search of a small acoustic feature set for emotion recognition is an important issue. Looking towards such an optimal set faces three main challenges. First of all, such a set must be robust to real-life applications, in other words, the search has to be conducted on real-life corpora. Next, emotion models are evaluated with an optimal set of acoustic features, simultaneously with parameters optimization. Finally, the result of feature selection must be evaluated in cross-corpus condition. The goal of the present study is to select the most consensual acoustic family for valence recognition via dependent (Gaussian Mixture Models and the Bhattacharyya distance) and independent (Information Gain) non-classification based feature ranking and cross-corpus experiments, and to optimize emotional models simultaneously.

Reducing the number of features goes in pair with optimizing model parameters (here SVM parameters). Experiments carried on randomly selected features from two acoustic feature sets (Li-174 and Os-384) show that a feature space reduction is needed to avoid over-fitting. A Grid search tends to find non-standard values with small feature sets. In order to ensure more stability of the SVM parameters, the authors propose a multi-corpus optimization method which consists in finding the most frequent optimized $(C; \gamma)$ values found on different corpora and acoustic feature subsets. Reducing the number of features implies to down perform the auto-coherence results but this reduction improves cross-corpus recognition rates. The results show that feature ranking and distance between GMMs are interesting methods to select acoustic families taking into account the mutual dependency of the features within the family. However, best individual feature or best family selected with the Information Gain method (Energy bands family) or best family selected with the G+B method (Spectral family) do not yield relevant results in cross-corpus experiments. Promising results have been obtained with a reduced set of 24 Voiced Cepstral coefficients while this family was ranked in 2nd (with IG) and 4th (with G+B) positions with both ranking methods. These results outperform the results obtained with the standard OpenSmile.

The present paper emphasizes many perspectives and further works. The results presented here show some interesting trends, that could be validated on larger real-life emotional corpora. Nonetheless, collection of more recorded spontaneous data is still a challenge in the emotion recognition field. In addition, further work is needed in selecting features (from different acoustic sets), improving modeling techniques and feature extraction and optimizing the parameters of the models simultaneously in the field of affective computing, especially for real-time recognition applications in the wild.

## Acknowledgement

This work was partially financed by the French projects: FUI ROMEO and BPI ROMEO2. The authors thank coders and co-workers who participated in elaborating protocols and annotating emotional states.

## References

[1] L. Devillers, L. Vidrascu, and O. Layachi, *Automatic detection of emotion from vocal expression*. Oxford University Press., 2010, ch. A blueprint for an affectively competent agent, Cross-fertilization between Emotion Psychology, Affective Neuroscience, and Affective Computing.

[2] L. Devillers and L. Vidrascu, "Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs," in *Interspeech*, Pittsburgh, PA, USA, 2006.

[3] A. Batliner, S. Steidl, B. Schuller, D. Seppi, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, V. Aharonson, L. Kessous, and N. Amir, "Whodunnit searching for the most important feature types signalling emotion-related user states in speech," *Computer Speech and Language (CSL), Special Issue on Affective Speech in real-life interactions*, vol. 25, Issue 1, pp. 4–28, 2011.

[4] C.-L. Huang and C.-J. Wang, "A ga-based feature selection and parameters optimization for support vector machines," *Expert Systems with Applications*, vol. 31, pp. 231–240, 2006.

[5] K. R. Scherer, "Vocal communication of emotion: a review of research paradigms," *Speech Communication*, vol. 40 (1-2), pp. 227–256, 2003.

[6] ——, *What are emotions ? and how can they be measured ?* SAGE Publications, 2005, ch. Social Science Information, vol 44(4), pp. 695–729.

[7] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *Journal of Research in Personality*, vol. 11, pp. 273–294, 1977.

[8] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, J. claude Martin, L. Devillers, S. Abrilian, A. Batliner, N. Amir, and K. Karpouzis, "The humaine database : Addressing the collection and annotation of naturalistic and. in induced emotional data," *Lecture Notes in Computer Science, Affective Computing and Intelligent Interaction*, vol. 4638, pp. 488–500, 2007.

[9] E. Douglas-Cowie, C. Cox, J.-C. Martin, L. Devillers, R. Cowie, I. Sneddon, M. McRorie, C. Pelachaud, C. Peters, O. Lowry, and et al., "The humaine database," *Cognitive Technologies*, vol. Emotion-Oriented Systems, Part 3, pp. 243–284, 2011.

[10] I. S. Engberg, A. V. Hansen, O. Andersen, and P. Dalsgaard, "Design, recording and verification of a danish emotional speech database," in *Eurospeech*, Rhodes, Greece, 1997, pp. 1695–1698.

[11] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. WeissI, "A database of german emotionnal speech," in *Interspeech*, Lisbon, Portugal, 2005, pp. 1517–1520.

[12] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge," *Speech Communication, Special Issue on "Sensing Emotion and Affect - Facing Realism in Speech Processing"*, vol. 53 (9/10), pp. 1062–1087, 2011.

[13] A. Batliner, S. Steidl, and E. Nöth, "Laryngealizations and emotions: How many babushkas?" in *Proc. Internat. Workshop on Paralinguistic Speech - between Models and Data (ParaLing' 07)*, Saarbrucken, Germany, 2007, pp. 17–22.

[14] L. Devillers, L. Vidrascu, and L. Lamel, "Challenges in real-life emotion annotation and machine leraning based detection," *Journal of Neural Networks, Special Issue on Emotion and Brain*, vol. 18 (4), pp. 407–422, 2005.

[15] A. Batliner, B. Schuller, D. Seppi, S. Steidl, L. Devillers, L. Vidrascu, T. Vogt, V. Aharonson, and N. Amir, *The automatic recognition of emotions in speech*. Springer, 2011, ch. Cognitive Technologies, pp. 71–99.

[16] E. Mower, A. Metallinou, C.-C. Lee, A. Kazemzadeh, C. Busso, S. Lee, and S. Narayanan, "Interpreting ambiguous emotional expressions," in *ACII*, vol. 978 (1), Amsterdam, The Netherlands, 2009, pp. 4244–4799.

[17] A. Batliner, C. Hacker, S. Steidl, E. Nöth, S. D'Arcy, M. Russell, and M. Wong, "«you stupid tin box» - children interacting with the aibo robot: A cross-linguistic emotional speech corpus," in *LREC*, Lisbon, Portugal, 2004, pp. 171–174.

[18] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schröder, "The semaine database: annotated multimodal records of emotionally coloured conversations between a person and a limited agent," *IEEE Transactions on Affective Computing*, vol. 3, Issue 1, pp. 5–17, 2012.

[19] J. G. Han, E. Gilmartin, C. D. Looze, B. Vaughan, and N. Campbell, "Speech & multimodal resources: The herme database of spontaneous multimodal human-robot dialogues," in *LREC*, Istanbul, Turkey, 2012.

[20] M. Tahon, A. Delaborde, and L. Devillers, "Real-life emotion detection from speech in human-robot interaction: Experiments across diverse corpora with child and adult voices," in *Interspeech*, Firenze, Italia, 2011.

[21] A. Delaborde, M. Tahon, C. Barras, and L. Devillers, "A wizard-of-oz game for collecting emotional audio data in a children-robot interaction," in *AFFINE'09*, Boston, MA, U.S.A., 2009.

[22] ——, "Affective links in a child-robot interaction," in *LREC*, Valetta, Malta, 2010.

[23] A. Batliner, D. Seppi, S. Steidl, and B. Schuller, "Segmenting into adequate units for automatic recognition of emotion-related episodes: a speech-based approach," *Advances in Human-Computer Interaction*, vol. 2010 (article ID: 782802), pp. 1–15, 2010.

[24] P. Ekman, *Handbook of cognition and emotion*. U.K.: Wiley, 1999, ch. Basic emotion.

[25] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. André, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, and K. Truong, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *Transaction on Affective Computing (IEEE)*, vol. x, pp. 1–14, 2015.

[26] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. S. Narayanan, "Paralinguistics in speech and language, state-of-the-art and the challenge," *Computer Speech and Language (CSL), Special Issue on Paralinguistics in Naturalistic Speech and Language*, vol. 27, Issue 1, pp. 4–39, 2013.

[27] A. Batliner, S. Steidl, , B. Schuller, D. Seppi, K. Laskowski, T. Vogt, L. Devillers, L. Vidrascu, N. Amir, loic Kessous, and V. Aharonson, "Ceices : Combining efforts for improving automatic classification of emotional user states: a ń forced co-operation ż initiative," in *Language and Technologies Conference*, Slovenia, 2006, pp. 240–245.

[28] C. Clavel, I. Vasilescu, and L. Devillers, "Fiction supports for realistic portrayals of fear-type emotional manifestations," *Computer Speech and Language (CSL), Special Issue on Affective Speech in real-life interactions*, vol. 25, pp. 63–83, 2011.

[29] B. Schuller, A. Batliner, D. Steppi, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson, "The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals." in *Interspeech*, Antwerp, Belgique, August 2007.

[30] C. Lee and S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Transaction on Speech and Audio Processing*, vol. 13 (2), pp. 293–303, 2005.

[31] T. Polzehl, A. Schmitt, and F. Metze, "Approaching multi-lingual emotion recognition from speech - on language dependency of acoustic/prosodic features for anger recognition," in *Speech Prosody*, Chicago, USA, May 2010.

[32] T. Kostoulas, T. Ganchev, A. Lazaridis, and N. Fakotakis, "Enhancing emotion recognition from speech through feature selection," in *International Conference on Text, Speech and Dialogue (TSD)*, 2010, pp. 338–344.

[33] M. Brendel, R. Zaccarelli, and L. Devillers, "A quick sequential forward floating feature selection algorithm for emotion detection from speech." in *Interspeech*, Makuhari, Chiba, Japan, September 26-30 2010.

[34] H. Altun and G. Polat, "Boosting selection of speech related features to improve performance of multi-class svms in emotion detection," *Expert Systems with Applications*, vol. 36, pp. 8197–8203, 2009.

[35] L. Meier, S. V. D. Geer, and P. Bühlmann, "The group lasso for logistic regression." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 70.1, pp. 53–71, 2008.

[36] D. Wu, T. D. Parsons, and S. Narayanan, "Acoustic feature analysis in speech emotion primitives estimation," in *Interspeech*, Makuhari, Chiba, Japan, 2010.

[37] J. Rong, G. Li, and Y.-P. P. Chen, "Acoustic feature selection for automatic emotion recognition from speech," *Information Processing and Management*, vol. 45, pp. 315–328, 2009.

[38] J. Tao and Y. Kang, "Features importance analysis for emotional speech classification," in *ACII*, vol. 13784/2005. Lecture Notes in Computer Sciences, 2005, pp. 449–457.

[39] B. Schuller, R. Zaccarelli, N. Rollet, and L. Devillers, "Cinemo - a french spoken language resource for complex emotions: facts and baselines," in *LREC*, Valetta, Malta, 2010.

[40] B. Schuller, B. Vlasenko, F. Eyben, M. Wöllmer, A. Stühlsatz, A. Wendemuth, and G. Rigoll, "Cross-corpus acoustic emotion recognition: variances and strategies," *Transaction on Affective Computing (IEEE)*, vol. 1, Issue 2, pp. 119–131, 2010.

[41] F. Eyben, A. Batliner, B. Schuller, D. Seppi, and S. Steidl, "Cross-corpus classification of realistic emotions: some pilot experiments," in *LREC, Workshop on EMOTION : Corpora for Research on Emotion and Affect*. Valetta, Malta: ELRA, 2010, pp. 77–82.

[42] Z. Zhang, F. Weninger, M. Wöllmer, and B. Schuller, "Unsupervised learning in cross-corpus acoustic emotion recognition," in *ASRU*, Honolulu, Hawaï, December 2011.

[43] E. Marchi, A. Batliner, and B. Schuller, "Speech, emotion, age, language, task and typicality: trying to disentangle performance and future relevance," in *Workshop on Wide Spectrum Social Signal Processing (ASE/IEEE International Conference on Social Computing)*, Amsterdam, Netherlands, 2012.

[44] B. Schuller, Z. Zhang, F. Weninger, and G. Rigoll, "Using multiple databases for training emotion recognition: to unite or to vote ?" in *Interspeech*, Florence, Italy, August 2011.

[45] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *Interspeech*, Brighton, U.K., 2009.

[46] B. Schuller, Z. Zhang, F. Weninger, and G. Rigoll, "Selecting training data for cross-corpus speech emotion recognition: prototypicality vs. generalization," in *AVIOS Speech Processing*, Tel-Aviv, Israël, 2011.

[47] S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan, "An acoustic study of emotions expressed in speech," in *ICSLP*, Jeju Island, Korea, 2004.

[48] L. Devillers, C. Vaudable, and C. Chastagnol, "Real-life emotion-related states detection in call centers: a cross-corpora study," in *Interspeech*, 26-30 September 2010, Makuhari, Chiba, Japan INTERSPEECH 2010 Makuhari, Chiba, Japan, 2010, pp. 2350–2353.

[49] C. Busso, S. Mariooryad, A. Metallinou, and S. Narayanan, "Iterative feature normalization scheme for automatic emotion detection from speech," *Transaction on Affective Computing (IEEE)*, vol. 4, no. 4, pp. 386–397, 2012.

[50] C. Chastagnol and L. Devillers, "Personality traits detection using a parallelized modified sffs algorithm," in *Interspeech*, Portland, Oregon, USA, 2012.

[51] M. Tahon and L. Devillers, "Acoustic measures characterizing anger across corpora collected in artificial or natural context," in *Speech Prosody*, Chicago, USA, 2010.

[52] K. W. Godin and J. H. Hansen, "Analysis of the effects of physical task stress on the speech signal," *J. Acoustic. Soc. Am.*, vol. 130 (6), pp. 3992–3998, 2011.

[53] B. Mak and E. Barnard, "Phone clustering using the battacharyya distance," in *ICSLP*, vol. 4, Philadelphia, PA, USA, October 1996.

[54] G. Sfikas, C. Constantinopoulos, A.Likas, and N. Galatsanos, "An analytic distance metric for gaussian mixture models with application in image retrieval," in *ICANN'05 Proceedings of the 15th international conference on Artificial neural networks: formal models and their applications*, vol. Volume Part II. Springer-

Verlag Berlin, Heidelberg, 2005, pp. 835–840. [Online]. Available: http://link.springer.com/chapter/10.1007%2F11550907_132#page-1
[55] M. Tahon, A. Delaborde, C. Barras, and L. Devillers, "A corpus for identification of speakers and their emotions," in *LREC*, Valetta, Malta, 2010.
[56] A. Delaborde and L. Devillers, "Use of nonverbal speech cues in social interaction between human and robot: Emotional and interactional markers," in *International Workshop on Affective Interaction in Natural Environements (AFFINE)*, Firenze, Italy, 2010.
[57] ——, "Impact of the social behaviours of the robot on the user's emotions: Importance of the task and the subject's age," in *Workshop on Affect, Compagnons Artificiels, Interaction*, Grenoble, France, 2012.
[58] L. Devillers and J.-C. Martin, "Coding emotional events in audiovisual corpora," in *LREC*, Marrakech, Morocco, 2008.
[59] C. Barras, E. Geoffrois, Z. Wu, and M. Liberman, "Transcriber: development and use of a tool assisting speech corpora production," *Speech Communication*, vol. 33(1), pp. 5–22, 2000.
[60] K. R. Scherer, *Affect Bursts*. Hillsdale, NJ: Lawrence Erlbaum, 1994, ch. Emotions, pp. 161–193.
[61] M. Tahon, G. Degottex, and L. Devillers, "Usual voice quality features for emotionnal valence detection," in *Speech Prosody*, Shanghai, China, 2012.
[62] M. Brendel, R. Zaccarelli, and L. Devillers, "Building a system for emotions detection from speech to control an affective avatar," in *LREC*, Valetta, Malta, 2010.
[63] C. Clavel, I. Vasilescu, L. Devillers, G. Richard, and T. Ehrette, "Fear-type emotion recognition for future audio-based surveillance systems," *Speech Communication*, vol. vol. 50, Issue 6, pp. 487–503, 2008.
[64] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the cuidado project," *Ircam*, 2004.
[65] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," *Institute of Phonetics Sciences, University of Amsterdam*, vol. 17, pp. 97–110, 1993.
[66] A. Röbel and X. Rodet, "Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation," in *Digital Audio Effects (DAFx'05)*, Madrid, Spain, September 20-22 2005.
[67] F. Ringeval and M. Chetouani, "Hilbert-huang transform for non-linear characterization of speech rythm," in *Workshop on Non Linear Speech Processing (ISCA)*, 2009.
[68] R. Ruiz, P. P. de Hugues, and C. Legros, "Advanced voice analysis of pilots to detect fatigue and sleep inertia," *Acta Acustica United with Acustica*, vol. 96, No.3, pp. 567–579, 2010.
[69] Z. Xiao, E. Dellandrea, W. Dou, and L. Chen, "A dimensional emotion model driven multi-stage classification of emotional speech," *Research report RR-LIRIS-2007-033*, 2007.
[70] L. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc, 1993.
[71] P. Dumouchel, N. Dehak, Y. Attabi, R. Dehak, and N. Boufaden, "Cepstral and long-term features for emotion recognition," in *Interspeech*, Brighton, U.K., 2009, pp. 344–347.
[72] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "The interspeech 2010 paralinguistic challenge," in *Interspeech*, Makuhari, Chiba, Japan, 26 - 30 sept 2010, pp. 2830–2833.
[73] B. Vlasenko, D. Prylipko, D. Philippou-Hübner, and A. Wendemuth, "Vowels formants analysis allows straightforward detection of high arousal acted and spontaneous emotions," in *Interspeech*, 2011.
[74] C. M. Lee, S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan, "Emotion recognition based on phonem classes," in *ICSLP*, Jeju Island, Korea, 2004, pp. 889–892.
[75] G. Beller, N. Obin, and X. Rodet, "Articulation degree as a prosodic dimension of expressive speech," in *Speech Prosody*, Campinas, Brasil, 2008, pp. 681–684.
[76] U. M. Fayyad and K. B. Irani, "Muli-interval discretization of continuous-valued attributes for classification learning," in *International Joint Conference in Artificial Interlligence*, 1993, pp. 1022–1027.
[77] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in *International Conference on Machine Learning*, 1997, pp. 412–420.
[78] G. B. Coleman and H. C. Andrews, "Image segmentation by clustering," *IEEE*, vol. 67, No. 5, pp. 773–785, 1979.
[79] F. Alam and G. Riccardi, "Comparative study of speaker personality traits recognition in conversational and broadcast news speech," in *Interspeech*, Lyon, France, August 2013.
[80] G. Chollet and C. Montacie, "Evaluating speech recognizers and databases," *Recent Advances in Speech Understanding and Dialog Systems, NATO ASI F: Computer and Systems Sciences*, vol. 46, pp. 345–348, 1988.
[81] L. Devillers, M. Tahon, M. Sehili, and A. Delaborde, "Inference of human beings' emotional states from speech in human-robot interactions," *International Journal of Social Robotics, Special Issue on Developmental Social Robotics*, vol. 7, pp. 451–463, 2014.
[82] S. J. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book Version 3.4*. Cambridge University Press, 2006.

**Dr. Marie Tahon** graduated in Engineering from the Ecole Centrale de Lyon (France) in 2007 and received the M.S. degree in mechanics, energetics, civil engineering and acoustics from the Ecole Centrale de Lyon, in 2007. She received the Ph.D. degree in informatics and signal processing from the University of Paris-Sud (Orsay, France) in 2012. She has been with the LIMSI-CNRS in the team "Affective and Social Dimensions of Spoken Interactions" (2009-2012 and 2014-2015). She has been a Teaching and Research Assistant in acoustics with the Structural Mechanics and Coupled Systems Laboratory (LMSSC), Conservatoire National des Arts et Métiers (Paris, France) (2012-2014). She is currently with the IRISA in the team "Expression". Her research interests concern automatic speech processing, i.e. automatic acoustic features extraction for both emotion recognition and speaker recognition. She is a member of the French Association of Spoken Communication (AFCP), of the French Acoustic Association (SFA) and of the Workgoup on Affects, Artificial Companions and Interactions (GT-ACAI).

**Prof. Laurence Devillers** is Professor of affective computing at Paris-Sorbonne IV University. She does her research at LIMSI-CNRS and heads the team on "Affective and Social Dimensions of Spoken Interactions" (https://www.limsi.fr/en/research/tlp/topics/topic2), working on machine analysis of human non-verbal behaviour including audio and multimodal analysis of affective states and social signals, and its applications to Human-Robot Interaction. She participates in BPI ROMEO2 project (2013-17) which has the main goal of building a social humanoid robot. She leads the European CHISTERA project JOKER (2013-16), JOKe and Empathy of a Robot/ECA: Towards social and affective relations with a robot. She also contributes to computer ethics, and is member of the working group on the ethics of the research in robotics of the CERNA. Prof. Devillers has co-authored more than 130 publications. She is a member of the board of AAAC (emotion-research.net) and Eurobotics, member of IEEE, ACL, ISCA and the French Association of Spoken Communication (AFCP).