Emotion Recognition in Never-Seen Languages Using a Novel Ensemble Method With Emotion Profiles

Enrique M. Albornoz, and Diego H. Milone, *Member, IEEE*

Abstract—Over the last years, researchers have addressed emotional state identification because it is an important issue to achieve more natural speech interactive systems. There are several theories that explain emotional expressiveness as a result of natural evolution, as a social construction, or a combination of both. In this work, we propose a novel system to model each language independently, preserving the cultural properties. In a second stage, we use the concept of universality of emotions to map and predict emotions in never-seen languages. Features and classifiers widely tested for similar tasks were used to set the baselines. We developed a novel ensemble classifier to deal with multiple languages and tested it on never-seen languages. Furthermore, this ensemble uses the Emotion Profiles technique in order to map features from diverse languages in a more tractable space. The experiments were performed in a language-independent scheme. Results show that the proposed model improves the baseline accuracy, whereas its modular design allows the incorporation of a new language without having to train the whole system.

Index Terms—Emotion Recognition, Ensemble Classifier, Emotion Profiles, Not-yet-encountered Languages.

1 Introduction

Interpersonal communication involves a lot of implicit and explicit information that can be present in speech, body language, facial expressions, and biosignals [1], [2], [3], [4]. Humans are very good at interpreting implicit information in these messages and they are able to arrive at diverse judgements about the messages and the speaker states. In the scientific community, the concept of *speaker state* is used in different scopes, where the word "state" can refer to emotional states, psychological states, intoxication or sleepiness degrees, or specific illness states. Over the last years, the recognition of diverse speaker states has become a multi-disciplinary research area that has drawn great interest [5], [6], [7], [8]. These issues play an important role in the improvement of human-machine interaction, security, and medical diagnosis, among others. These applications, as commercial products, would have to deal with some current challenges, such as coupling of tasks, continuous modelling, robustness, more realism, cross-corpus, and the ability to operate using never-seen languages [9].

Emotions have been long debated by psychologists, confronting those who propose universality versus those who argue that the expression of emotions vary by culture [10], [11], [12]. Although the theories seem to be in opposition, recent researches attempt to consider both universality and cultural variation [13], [14]. Emotion perception is the ability to recognise emotions in faces, voices, and cultural artifacts (such pictures or music) and the cross-cultural studies are mainly supported by analysis of facial expression and have been supported less often by research in vocal expressions [10], [15]. As can be noted in the current literature, this challenge have not been solved in the psychological field yet and its application in automated systems is further novel. Recently, Pell et al. [14] have studied how monolingual speakers are able to recognise basic emotions from nonsense utterances produced in their native language and in foreign languages. They argue that the people ability to understand emotions in speech is partially independent of linguistic ability and involves universal principles, although this ability is also shaped by linguistic and cultural variables. In [13], the mechanisms underlying the human perception of emotional expressions are investigated using the cross-modal analysis of realistic emotional stimuli. The authors analysed the effect of culture and language specificity on emotion recognition, and studied how the subjects familiarity and exposition to the language and culture influences their perception. Meanwhile, Argstatter [16] evaluates if perception of basic emotions in music is culture-specific or multicultural. His results give some evidence of pan-cultural emotional sentience in music. However, some cultural, emotional and item-specific differences have arisen in emotion recognition.

As mentioned in [17], it is possible to say that only a few recent studies address cross-corpus emotion recognition and, furthermore, issues like emotion recognition on never-seen languages need to be investigated. Moreover, the models should manage the specific information of each language and should propose techniques to map the information or to perform late fusion with specialised classifiers. Dealing with never-seen languages is not trivial but, as already stated, it is an desired and necessary functionality that the next generation systems should have. For example, we can consider a security system that evaluates the stress level of a speaker. It is very important that the system works

• *Enrique M. Albornoz and Diego H. Milone are affiliated to the Research Institute for Signals, Systems and Computational Intelligence, sinc(i), FICH-UNL/CONICET, Argentina. Univ. Nacional del Litoral CC 217, C. Universitaria, Paraje El Pozo, S3000 Santa Fe. Tel./Fax: +54 (342) 457-5233 ext 191; http://fich.unl.edu.ar/sinc*
E-mail: emalbornoz@sinc.unl.edu.ar, dmilone@sinc.unl.edu.ar

Manuscript received April 29, 2015; revised September 25, 2015.

with never-seen languages since training the system for all possible languages would not be feasible. Although the system might be able to identify the spoken language, either automatically, informed by user or using the area code in a telephonic system, it may not have the amount of information needed to create a language-dependent system. In addition, if the language is identified, it may be possible to adapt the system using information obtained from the more similar languages for this task. An approach to this case will be addressed in this work.

The use of speech signals is the most widespread in the context of emotion recognition, possibly because it carries more information than others (it could be debatable depending on what kind of information is being discussed) and to the earlier availability of the first databases for this task. Feature extraction of speech has been focused on different aspects: speech production, characteristics of speech signals, speech perception, etc. Most researchers have analysed the prosodic features and spectral information [18], [19], [20]. Some of the widely-known features used in literature are Mel-frequency cepstral coefficients (MFCC), linear prediction cepstral coefficients (LPCC), perceptual linear prediction coefficients (PLPC), and formant features [21], [22], [23]. With regard to classification, several standard techniques have been explored for emotion recognition: Gaussian Mixture Models (GMM), Hidden Markov Models (HMM), Multilayer Perceptron (MLP), Support Vector Machines (SVM), k-nearest neighbour (k-NN), and Bayesian classifiers [22], [24], [25], [26]. At present, the combination of standard methods, such as fusion, ensemble or hierarchical classifiers, has become the focus of state-of-the-art studies [27], [28], [29], [30].

As Schuller and Weninger [9] mentioned, some currently dominant trends might characterise the next decade of research in computational paralinguistics. Despite the recent developments in this field, there are still some *black spots* in literature. One of these *trend topics* is to overcome cross-language and cross-cultural barriers. Usually, experiments to classify emotions are performed with cross-validation using one corpus. For this scheme, called "within corpus", the community recommends the Leave-One-Subject-Out (LOSO) cross-validation as the most suitable technique to ensure speaker independence [31]. These approaches (using one corpus) lead to highly adapted models that have not enough generalisation capability for other corpora [17].

Due to the lack of specific studies on recognition of emotion in unseen languages, the background for similar tasks will be reviewed. Multi-corpus analyses have been explored in [32], where the focus is on testing some techniques for feature extraction or for classification using several databases in an isolated way. The cross-corpus evaluation for emotion recognition is appropriate to judge the generalisation ability of the models [17]. In addition, a commercial product would frequently have to deal with this kind of testing conditions. Another advantage of cross-corpus evaluation is the easy reproducibility of the results because the partitions are well defined in the task itself. In one corpus emotion recognition, this issue has been addressed and many emotional databases include explicit partitions for training and testing (e.g., FAU Aibo, RECOLA). Although cross-corpus is an interesting topic, very few studies use one corpus for training and a different one for testing [33], [34]. Schuller et al. [17] proposed a model to evaluate cross-corpora with six corpora using a well-known set of features and support vector machines. They used categorical and dimensional labels in diverse experiments with different numbers of classes. In each case, the training sets were composed of one corpus or a combination of corpora, while a different corpus was used for testing. In any case, they showed a poor performance in cross-corpus recognition, which was partially improved by using corpus or speaker normalisation. In the same line of thought, some researchers have worked con cross-language including less researched languages of more distant language families such as Burmese, Romanian or Turkish [35]. Eyben et al. [36] used four corpora to test cross-corpus emotion recognition. Categorical and dimensional labels are mapped onto a representation of valence with the three classes: positive, neutral, and negative. The exploration to find a set of generic and corpus-independent acoustic features was partially successful as stated in [36]. Common techniques of feature selection show low cross-data generalisation and do not reduce over-fitting. In a similar way, Schuller et al. [34] presented an interesting study using a combination of databases with a unified labelling scheme (dimensional labels: arousal and valence). They proposed data agglomeration and voting to improve the performance of classifiers. However, the unification of labels involves a loss of information because this mapping cannot provide the same meaning expressed by the original human labellers. An alternative could be to guide the selection of features to those that optimise the classification in the multitasking scheme (multiple languages). This idea was recently presented to recognise emotions from singing and speaking [37].

In the context of emotion recognition systems in never-seen languages, it is important to work on the hard classification idea (prototypical classification) in order to take advantage of the emotional information of the available languages in the system modelling. Unlike traditional classification problems where pattern labels must be definitely correct, in emotion recognition this is not usually guaranteed. Corpora are commonly labelled by a group of humans who do not know which emotion was really expressed by the speaker. Consequently, the labels for an utterance do not agree on one common class and similar classes are regularly confused. Finally, the resulting labels could be split in prototypical emotions, when the utterances are consistently classified by a set of human evaluators, and in non-prototypical emotions, when they are not recognised consistently [38]. As can be seen, this situation becomes even more complicated when trying to model the prototypical emotions using several languages. A way to address this in a standard emotion classifier is by using different weights for those classification mistakes that also take place in human labelling [39]. The ambiguity in the labelling made by humans, evidences that one utterance could have non-prototypical emotional content ([38], [40]). It means, the utterance could have a relationship with several emotion classes, maintaining a certain degree of membership to each class. We believe that this idea can be very productive in a multiple-language scheme to predict emotions in never-seen languages because non-prototypical emotions are expected.

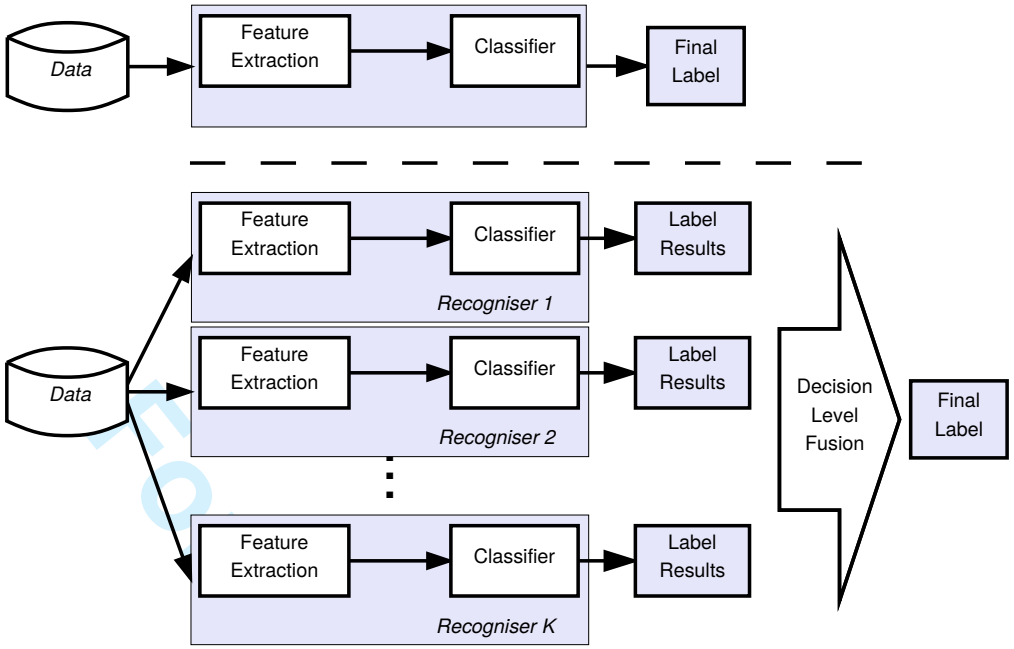

Figure 1. Graphical schemes of a standard classifier (top, SC) and an ensemble classifier (bottom, EC).

Mower et al. [38] suggest that classifying non-prototypical utterances using only models trained with prototypical utterances may be unfavourable. In this line of thought, the Emotion Profiles (EP) [41] and the Anchor Models (AM) [42] have been defined for emotion recognition. There are some works that explored the use of EP and AM for language and speaker recognition [43], [44] and music recommendation [45], and some recent works proposed these techniques for emotion recognition [46], [47], [48], [49], [50]. Cao et al. [51] presented a similar idea through a ranking approach where binary-ranking SVMs are trained for individual emotions (one-vs-all) and the predictions from all rankers are combined to perform a multi-class prediction. A recent work employed a weighted SVM method to demonstrate that the best accuracy is reached when prototypicality information is explicitly used [52].

In this work, we propose an emotion recognition model to deal with unseen language problems. Our model is designed in an ensemble framework where the diverse languages used for training are independently modelled and the results for the tested language are fused at decision level. The EP approach is used as a mapping method inside the model. The model is tested in a leave-one-language-out scheme using categorical labelled corpora. Furthermore, we present a discussion of a practicable application and the analysis of some similarities among languages that could indicate what language is more predictable from another one and what kind of languages could be used together to train a model.

In the next section, the proposed model is introduced. Section 3 deals with the experimental set-up and presents the implementation details. In addition, it explains the validation schemes and addresses a real case application. Results are presented and discussed in Section 4. Finally, conclusions and future works are presented in the last section.

2 Standard Classifiers and Proposed Model

The general schemes of traditional and state-of-the-art classifiers are introduced below and then a new classifier for emotion recognition in never-seen languages is presented.

Firstly, two well-known classifiers for emotion recognition are introduced (Fig. 1). The simplest one is the single standard classifier (*SC*), where a set of features is used to classify patterns among the different classes. The other one is the ensemble classifier (*EC*), where, keeping the same goal, a set of classifiers is trained and the final decision is a combination of predictions from these multiple classifiers. Each single classifier can have a specific set of features and configuration in order to achieve more versatility and confidence. Ensemble classifiers can be used in two basic ways: *classifier selection* or *classifier fusion* [53]. In the first one, each classifier is trained in a specific region of the pattern space and then its decision will be more important when a neighbour pattern is classified. In the other case, all classifiers are trained using the entire feature space and then the key is to combine them to achieve a lower error. In this work, we propose an ensemble scheme (bottom of Fig. 1) where each recogniser (feature extraction and classifier) is adapted and trained using one specific language. For the final decision, two strategies are used to combine ensemble member decisions: *majority voting* and *combining continuous outputs*. In majority voting the class C^* is the one that receives the highest number of votes, even when the sum does not exceed 50% of the votes [53].

$$C^* = \max_c \sum_{k=1}^{K} d_{k,c}(\mathbf{x}) \ , \qquad (1)$$

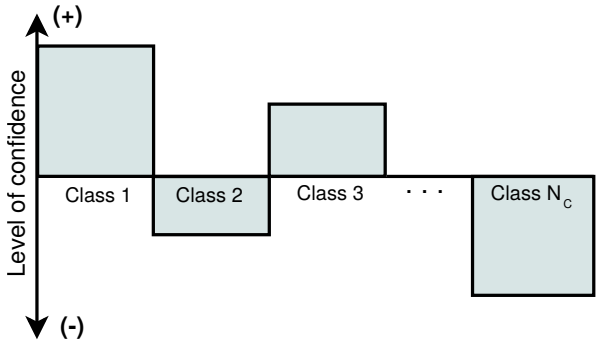

Figure 2. Schematic representation of an Emotional Profile: level of confidence for each class-specific binary classifier.

where $c \in [1 \dots N_c]$ and N_c is the number of classes, K is the number of classifiers, $d_{k,c}(\cdot) \in \{0, 1\}$ is a binary decision of the k^{th} classifier with respect to class c, and $\mathbf{x}$ is a feature vector.

Usually, it is possible to have classifiers that provide continuous output for the classes ($d_{k,c}(\cdot) \in [0, 1]$). This could be considered as the degrees of support given to the classes by the classifier, and sometimes it can also be understood as posterior probabilities. In order to apply the diverse combination rules, and for a better understanding, Kuncheva's decision profile matrix is defined as $KDP(\mathbf{x})$, where each row is the support of a classifier for all the classes and each column has the total support for a specific class. In this work, both binary and continuous outputs are used. Graphical schemes of Standard Classifier (SC) and ensemble classifier (EC) are showed in Figure 1.

2.1 Emotional Profile Classifier

The EP technique describes the confidence between every emotional label and the utterance; that is to say, the utterance is expressed in terms of the emotions present in that utterance [41]. Hence, in the case of a non-prototypical emotional expression, the emotional information is kept and it would be associated to the emotional labels with different degrees of membership. These models are implemented basically in two stages. In the first stage, the profiles are created by mapping the source features into a profile space, whereas, in the second stage, the profiles represent the inputs of a classifier that gives the final decision. The dimension of profiles is defined by the number of classes (N_c) and the values are similarity scores. These values are usually computed with N_c binary classifiers, where each one is pre-trained using one class versus all the rest (Fig. 2). The methods described below follow closely these ideas, extending them to different types of classifiers.

The implementation of a classifier based on Emotional Profile (*EPC*) could be explained as a two-step classifier. First, a set of N_c emotion-specific binary classifiers is created. Each classifier is trained to distinguish between class *class c* and *not class c*, and each feature set could be optimised. Then, the continuous outputs of the binary classifiers are used to create the EP vector that represents the confidence of each emotion-specific decision (Fig. 2). In the second stage, the EP vector is used as input of a multi-class classifier and, finally, a class label is assigned to the utterance. As it was discussed, the EP classifiers would have an inherent benefit when the utterances have non-prototypical content. Therefore, it would be appropriate to consider this type of classifiers in a multi-language scheme. While Ekman et al. [54] showed that humans can distinguish the basic emotions across cultures, the perceptual tests on different corpora indicate that non-prototypical content should be considered to get more useful information in automatic systems. Figure 3 shows a general diagram of a classifier based on EP.

2.2 Proposed Classifier For Emotions In Never-Seen Languages

In this novel task, we present a model that can be trained using information from a set of available languages and that can predict emotions in a never-seen language. We take the standard leave-one-out technique for emotion recognition and adapt it to define a leave-one-language-out scheme. A first strategy could be to leave one language out for testing while all the others are put together to train the system. To do this, a normalisation over all training data would be required, which leads to some language characteristics being lost at feature level. In a different way, our system is defined to model each single training language with a different set of features and classifier. In this first stage, it tries to model the language peculiarities and then to map the information into a common space using the Emotional Profiles. The aim is to reach a common space from the specific characteristics of each language. In the second stage, the EP are classified in the emotional labels using one classifier per language. This step is performed because categorical labels are expected as final output of the system. Finally, the categorical labels are combined in the decision-level fusion (late fusion) scheme to get the final result. It is expected that the combination of information obtained independently will be the best approach to the final hypothesis. In order to apply this fusion, (1) is used. In summary, our system allows extensive flexibility in choosing features and classifiers for each training language while it avoids the correlation at feature level that is produced when several languages are used together. Despite the more tedious learning process, the modular design is favourable for the isolated re-training of each language and it allows easily including a new language to the trained system. Figure 4 shows a graphical scheme of the proposed model called ensemble EP classifier (*EEPC*). In the test stage, the never-seen language k is evaluated in every module (trained for languages $1 \dots k-1$) and the final label is computed considering the $k-1$ categorical emotional labels.

3 Experiments

3.1 Emotional Speech Corpus

Publicly available speech data for different topics of research in paralinguistics is still sparse [9] and one of the biggest problems of cross-corpus emotion recognition is the mismatched acoustic conditions between training and test data [17]. Despite the large number of corpora available for the emotion recognition task, they usually have different

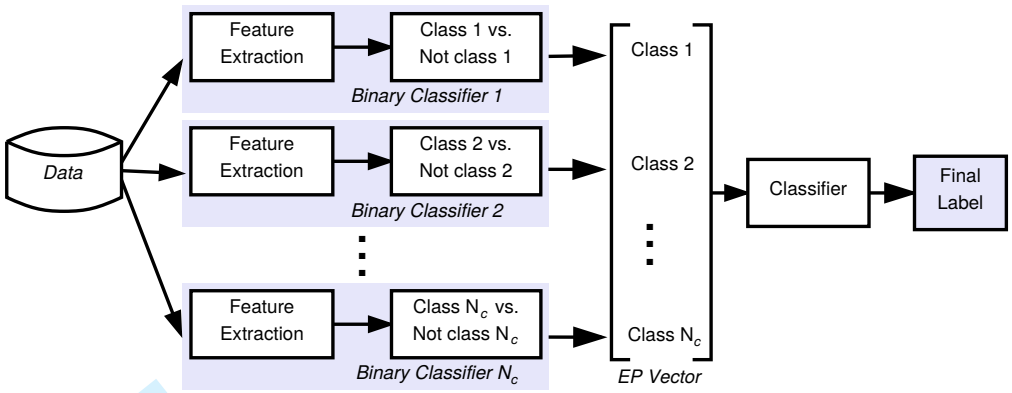

Figure 3. Diagram of an Emotional Profile Classifier (EPC).

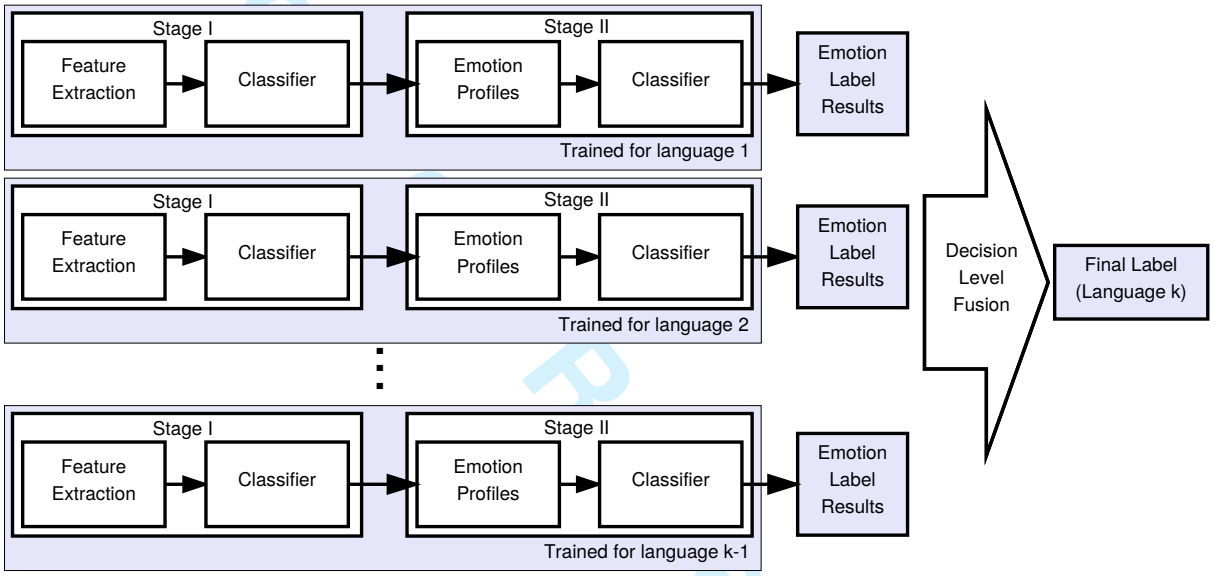

Figure 4. Graphical scheme of the proposed Ensemble Emotional Profile Classifier (EEPC).

recording conditions; are developed as acted, elicited or spontaneous; have different labels (categorical, dimensional, etc.); are expressed in different languages; or have an imbalance among classes. Furthermore, it is not trivial to find a database developed in several languages simultaneously [55]. In order to obtain performances influenced only by language characteristics, corpora recorded under the same conditions are required. Since our proposal aims to be an approach to a system that can model languages independently while providing language-independent results, we performed experiments using RML emotion database [56]. The corpus is freely accessible[1] and was used in several recent studies [57], [58], [59]. The video clips were recorded with a digital camera (using 30 FPS) in a bright and static environment, with a simple background. Each sample has a length of about 3-6 seconds with an initial silence, and the audio were recorded at a sampling rate of 22050 Hz. The corpus contains 720 audiovisual emotional utterances from 8 male subjects, speaking 6 languages and expressing 6 basic emotions: anger, disgust, fear, happiness, sadness, and surprise. The distribution of the languages and subjects is: English (7 subjects), Mandarin (4 subjects), Urdu (4 subjects), Punjabi (1 subject), Persian (3 subjects), and Italian (2 subjects). For each subject, the amount of utterances for each emotional class is balanced (about 5 sentences per emotion for each subject). Then, the distribution over languages is: Urdu $\sim 4 \times 5 \times 6 = 120$ utterances, Persian $\sim 3 \times 5 \times 6 = 90$ utterances, and so on. Each utterance has one categorical emotion label. The subjects had to express their emotions as naturally as possible, recalling experiences of their lives. A list of emotional sentences were used as reference only, then the subjects could express their emotions by using the same sentence structure, variations or different sentences according to their cultural background.

In addition, it is important to mention that this database allows evaluating the model with six categorical labels, different from cross-corpus approaches where these are mapped in some dimensional labels. Moreover, a normalisation across corpora ([17]) would not be necessary because the different languages have been recorded under the same conditions.

1. http://www.rml.ryerson.ca/rml-emotion-database.html.

3.2 Features Extraction

Although the database is audiovisual, in this work only speech is considered. A Wiener-based noise filter ([60]) was applied to the audio signals in order to reduce the noise present in the recordings. As all utterances have an initial silence, the noise could be easily modelled. Then, the endpoints of speech utterances were computed using a voice activity detector based on Rabiner and Schafer method [61]. Audio segments in which the power of signal was lower than 1% of the maximum power of signal were discarded.

The features used in this work included the prosodic and spectral ones [21], [22]. The prosodic features in emotion recognition have already been studied and discussed extensively [18], [62], [63]. The toolbox provided by Giannakopoulos and Pikrakis [61][2] was used to compute the prosodic features: zero crossing rate, energy, energy entropy, and fundamental frequency. The spectral features include spectral entropy and MFCC, widely known in emotion recognition [21], [22], [28]. For each emotional utterance, the first 13 MFCC were calculated within Hamming windows of 25 ms with a 25 ms frame shift (using AALC toolbox). We also considered the mean of the log-spectrum (MLS) coefficients, defined as

$$S(k) = \frac{1}{N}\sum_{n=1}^{N} \log |v(n,k)|, \qquad (2)$$

where k is a frequency band, N is the number of frames in the utterance, and $v(n,k)$ is the discrete Fourier transform of the signal in frame n. These were computed using spectrograms from non-overlapped Hamming windows of 25 ms. The first 30 MLS coefficients, corresponding to lower frequencies ($0 - 1200$ Hz), were considered because they have the most useful information [29]. In addition, a novel set of features based on an auditory spectrogram is used for emotion recognition. Yang et al. [64] proposed a model based on neurophysiological investigations at various stages of the auditory system. This model consists of two stages. The first one allows obtaining an early auditory spectrogram of the temporal signal at the auditory nerve fibres level. The second stage mimics a model of primary auditory cortex in mammalians to process the spectrogram. The first part of the model is composed of a bank of cochlear overlapping filters with centre frequencies that are uniformly distributed along a logarithmic frequency axis. In five steps, feature V5 results the short-time average computed on the positively-valued derivatives, which are calculated on the amplitudes of extrema points of each cochlear filter output. This process provides 128 coefficients representing the range of 0 to 4000 Hz, not equally distributed in frequencies (for example, the first 71 coefficients correspond to the $[0 - 1200]$ Hz interval). The quantity and frequency distribution of the filters proved to be satisfactory for discriminating important acoustic clues and for appropriately reconstructing speech signals [65].

The mean of the log-spectrum using the auditory spectrogram (MLSa) is defined as

$$S_a(k) = \frac{1}{N}\sum_{n=1}^{N} \log |a(n,k)|, \qquad (3)$$

where k is a frequency band, N is the number of frames in the utterance and $a(n,k)$ is the k-th coefficient obtained by applying the auditory filter bank to the signal in frame n. The MLSa was computed using auditory spectrograms calculated for windows of 25 ms without overlapping. In order to obtain the representation of sound in the auditory model, a Matlab implementation of the Neural System Lab auditory model was used[3]. As in the MLS case, only the $[0 - 1200]$ Hz range was considered. All features were computed at frame level, and then the mean and standard deviation of all features over the whole utterance was calculated. Finally, the 238 features were arranged in a vector for each utterance. A feature selection approach is carried out in each classifier and it will be explained in the next section.

3.3 Classifiers

To perform the experiments, Support Vector Machine was chosen as a classifier since it is a supervised learning method widely used in this field of study. In order to apply this, the LIBSVM library[4] was used. In each classification task, several features sets were created using the F-Score measure [66]. F-score is a simple method which measures the discriminative capacity of two sets of real numbers. Then, given the feature vectors in $\mathbb{R}^N$, the N features are ranked depending on their discriminative capacity [67], and the feature sets are formed using the k best features. For every feature set, SVM parameters were explored in order to create the best classification model. Radial basis function kernels are used in the SVM models and their accuracies were computed using a 5-fold cross-validation scheme, considering only the training data. As a result, we obtained the weighted accuracy and the best parameters for each feature set. In a second step, a new SVM model was trained with the whole training data for the task, using the settings that achieved the best accuracy in the exploration step.

It is noteworthy that all experiments were also performed using multilayer perceptrons; however, the results are not presented here because the SVMs were better even in the case of multi-class classifiers.

3.4 Validation

This section introduces the different validation schemes used in the experiments. It is important to mention that, for all the experiments, all the coefficients in vectors were normalised. For that purpose, maximum and minimum values (for each dimension) from the training set were extracted, and then the training and test vectors were normalised using these values.

Within corpus scheme

For each language, we compute the within-language recognition accuracy by 4-fold stratified cross-validation (SCV). This preliminary classification allows setting the first baselines, and these results would represent the maximum achievable score if the system is trained and tested using the same language. Our objective is to predict emotions in never-seen languages, so these results will only be the first guidelines.

2. AALC toolbox

3. Neural Systems Lab., Institutes for Systems Research, UMCP. http://www.isr.umd.edu/Labs/NSL/

4. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm/

Cross-language scheme

In our experiment, we considered the *leave-one-language-out* cross-validation technique to obtain a performance estimation on never-seen languages. In this line of thought, we introduce two cross-language schemes (considering N languages):

- 1**-vs-**1**:** in this case, two languages are available, so one language is used for training and the other one, for testing.
- $(N-1)$**-vs-**1**:** in this case, one language is used for testing, while the remaining languages are available for training. In this situation, we propose using the training data in two ways: as *mixed language set* and as *late fusion of the classifiers modelled for each language*. In the first case, all the training languages are grouped together in one set to train the system, while, in the second one, language-specific classifiers are modelled and the results are combined to obtain the final decision.

A simulated case of real application

All the mentioned validation methods lead to emotion recognition in a never-seen language; then the system could be considered blind with respect to the tested language. In this section, we propose a simulated case of real application where the tested language is known, but there is no available data to train the model. For example, an application that works on telephony could identify the language by using the area code or some automatic language recognition system [43]. If we think of an application for a specific task, it would be very usual not to have enough data to train the system for all possible languages. However, it would be possible to know which languages are more suitable for predicting emotions in other languages. Following this line of thought, the system should be capable of managing different languages and modelling each one independently, and then knowing the tested language to arrange the system appropriately to achieve an optimal result. Our proposed classifiers allow this flexibility and some preliminary results are discussed in the next section.

4 Results and Discussion

The first classification results are achieved by training and testing on each language in isolation, considering the six emotional classes. Even though our objective is to predict emotions on never-seen languages, these results give us an idea about the performance that can be obtained in the best situation (appropriate data is available for training). Then, 4-fold cross-validation per language were used and the average accuracies are shown in Table 1. Each SVM classifier was optimised using 5-fold CV on its training partition, as was mentioned in Section 3.3. On average, performance is 70.58% using this classification scheme (language dependent).

In Table 2, we present the results obtained for the six emotional classes using one language to train the system and another one for testing. The languages used for training are in rows, while the columns show the languages used for testing. This scheme supports the first approach of cross-language scheme (1-vs-1) to emotion recognition in never-seen languages. The results show what could be expected from a system trained with a different language, reaching 45% on average. Summarising the results presented, it is possible to say that this is the reference score for emotion recognition in a never-seen language using a standard multi-class classifier. As 70.58% was the average score for a language-dependent scheme, this could be considered as the maximum achievable score for the never-seen language scheme under these experimental conditions.

Table 1
Results of Training and Testing on the Same Language.

Language	Mandarin	English	Italian	Persian	Punjabi	Urdu
Average accuracy	0.81	0.70	0.45	0.61	0.84	0.83

Table 2
Results of Using One Language for Training and a Different One for Testing.

Training languages	*Tested languages* Mandarin	English	Italian	Persian	Punjabi	Urdu
Mandarin		0.42	0.35	0.33	0.54	0.43
English	0.52		0.33	0.54	0.77	0.62
Italian	0.35	0.34		0.40	0.31	0.41
Persian	0.40	0.45	0.43		0.46	0.55
Punjabi	0.39	0.35	0.22	0.41		0.60
Urdu	0.56	0.43	0.31	0.49	0.81	

Table 3
Performance on Never-seen Languages, Trained With Mixed Languages.

Classifiers	*Tested languages* Mandarin	English	Italian	Persian	Punjabi	Urdu
SC	0.60	0.53	0.47	0.56	0.65	0.60
EPC	0.50	0.48	0.41	0.54	0.65	0.53

As we mentioned, there are some ways of combining the available languages in order to train the system, keeping one never-seen language for testing. Now we present the results obtained using mixed languages (set of $N-1$ languages) in the training sets. For these experiments, we implemented the two classifiers described in Section 2: *SC* in Fig. 1 and *EPC* in Fig. 3. The results are shown in Table 3, where the used classifiers are presented in the first column and the other columns show the performances on different never-seen languages. As can be seen in the table, the standard multi-class SVM performs better than the classifier implemented using EP. On average, *SC* reaches 56.8%, whereas *EPC* obtains 52.1%. These results could indicate that the EP models are not able to handle different emotion classes in several languages simultaneously. While emotions of the same class could share information among diverse languages, the arrangement of classes in the feature space would not be similar for different languages. Therefore, the binary division of the space proposed by EP models would

Table 4
Performance on Never-seen Languages, With Modular Training and Late Fusion.

Classifiers	*Tested languages*					
	Mandarin	English	Italian	Persian	Punjabi	Urdu
EC[#1]	0.53	0.44	0.29	0.52	0.81	0.65
EC[#2]	0.54	0.47	0.41	0.54	0.73	0.68
EEPC[#1]	0.48	0.40	0.35	0.54	0.77	0.62
EEPC[#2]	0.52	0.48	0.41	0.57	0.81	0.70

not be useful for this scheme (multiple mixed languages) and another scheme is required to develop a system that can handle multiple languages, preserving the EP advantages. These results show that classical schemes to process prototypical emotions seem not to perform well when the labels for an utterance do not agree on one common class. The Emotional Profiles is an appropriate technique to deal with these cases. Furthermore, utterances with ambiguous emotional contents would be more likely in schemes where emotion recognition is carried out on a language different from those used for training.

The last set of experiments is performed on classifiers with modular language modelling and late fusion of results. As we stated, our objective with this new scheme is to keep the characteristics of the language as much as possible. Table 4 shows the classification accuracy for *EC* and *EEPC*. For the combination of class labels, both results fusion techniques were considered: [#1] majority voting and [#2] combination of continuous output. The first column presents the classifiers and the remaining columns show the performance on each never-seen language. As can be observed in this table, all classifiers work well in this scheme and the combination of continuous output ([#2]) is always better than majority voting ([#1]), except for Punjabi. The proposed *EEPC* [#2] reaches a 58.1% on average, outperforming the *EC* [#2] in about 2%.[5] It is important to note that the *SC* method perform well, however, a direct comparison with ensemble methods could not be fair. The modular models have important advantages: they should not be completely trained if a language is added; each module can have a weight in order to improve a decision for a particular language; and train and test can be done in parallel in each module, among others.

As mentioned above, for a specific application, it is possible not to have enough data to train the system for a particular language. However, this unseen language could be known. In addition, it would be possible to think about getting an estimation of the performance on the unseen language using a system trained with only one different language. A possible result is that shown in Table 1. In this table, it can be seen which languages are better to predict other languages and, in this regard, a priority order could be established. Using such a priority order and the proposed modular model *EEPC*, it is possible to configure the system to use n known languages to predict emotions in the never-seen language. Then, for each unseen language, different

5. All these results are computed in terms of unweighted accuracy. We included the results in terms of precision, recall and F1-score as supplementary material.

Table 5
Performance on Never-seen Languages Combining the Best Predictor Languages.

# predictor languages	Fusion method	Average accuracy
all the available (5)	majority voting	0.53
	continuous output	**0.58**
the top 4	majority voting	0.55
	continuous output	**0.58**
the top 3	majority voting	0.55
	continuous output	**0.59**
the top 2	majority voting	0.53
	continuous output	**0.58**
the best (1)	majority voting	0.56
	continuous output	0.56

numbers of classifiers were used to train the system and the average results are presented in Table 5. In the first column, the amount of classifiers used to train the system is presented. The fusion methods are indicated in the second column, while the average accuracy is informed in the last one. The results obtained using only one language are the same for both fusion methods, as it was expected, while the results obtained using all available languages are the ones presented previously (see Table 4). It can be observed that combining continuous output gets always a better result and the combination of the three languages with higher priority produces the best average result. Although the differences between these averages are not significant enough to draw definitive conclusions, the variances tend to decrease when more language are included in the decision.

In a last analysis, we have considered a non-fixed number of languages to predict each language; that is, emotions in an unseen language were predicted using three languages, while four languages were used to predict emotions in another language. This can be seen as a weighted combination of several languages, in which the weights can be real or binaries. These decisions will be led by the individual performance on each language, and for this case (binary combination) the optimal numbers of languages used to predict emotions are the following: 5 for Persian and Urdu, 3 for English, 2 for Mandarin, and 1 for Italian and Punjabi. Then, if the best adapted systems are used to recognise emotions in the different languages, a 62.47% on average is reached. This result is promising because it is better than our blind approach (58.1%) and greater than the reference average score achieved using one language for training (45%). In addition, it is near the 70.58% obtained when the system is trained and tested using the same language (that is, the optimal condition).

The proposed approach to model the training languages independently has two main advantages with respect to mixed languages at data level. Firstly, there are no problems of imbalanced data between languages and the results are not biased in favour of a particular language. Secondly, if normalisation is required, this will be performed on each language and their parameters will not change if a new language is added. Furthermore, the proposed system has the ability to deal with the different training languages

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
E. M. Albornoz & D. H. Milone; "Emotion Recognition in Never-Seen Languages Using a Novel Ensemble Method With Emotion Profiles"
IEEE Transactions on Affective Computing, 2016.

as modules, which are trained in an isolated way, and the incorporation of a new module only requires a minor change in the fusion function.

5 Conclusions and Future Work

In this paper, we have presented a scheme to classify spoken emotions in never-seen languages using categorical labels. The proposed methods allow modelling each language by using specific features and classifiers. The final decisions in the never-seen language are computed as a fusion of decisions obtained from classifiers trained with known languages. We have shown that the proposed model is improved when an EP mapping is considered within each language. The results indicates that the models for dealing with non-prototypical emotions are useful in cross-language schemes. Regarding the objective evaluation, we have defined the average baseline (45%) on never-seen languages, using one language to train the system. Also we proposed the optimal reachable score in 70.58%, considering the same language for training and testing the system. Standard and EP classifiers were evaluated on never-seen languages using training sets with mixed languages. On the other hand, ensemble systems with late decision were proposed to model the training languages independently. The *EEPC* model includes an EP mapping and reaches the best performance (58.1%). We also analysed how this proposal could be used in a real application with better performance (62.47%). In future work, the EP models could be improved, for example, by using fusion of diverse classifiers or normalisation strategies for the different languages. Also, more features and statistical functions, such as those in openSMILE Emobase, will be considered. Furthermore, we will investigate the performance of the proposed model when each language-dependent component is trained using data from different corpora.

Acknowledgments

The authors would like to thank the *National Agency for Scientific and Technological Promotion* (ANPCyT) and *Universidad Nacional de Litoral* (with PACT 2011 #58, CAI+D 2011 #58-511), as well as the *National Scientific and Technical Research Council* (CONICET), from Argentina, for their support.

References

[1] G. Chanel, J. J. Kierkels, M. Soleymani, and T. Pun, "Short-term emotion assessment in a recall paradigm," *International Journal of Human-Computer Studies*, vol. 67, no. 8, pp. 607–627, 2009.

[2] D. Giakoumis, D. Tzovaras, and G. Hassapis, "Subject-dependent biosignal features for increased accuracy in psychological stress detection," *International Journal of Human-Computer Studies*, vol. 71, no. 4, pp. 425–439, 2013.

[3] K. Schindler, L. Van Gool, and B. de Gelder, "Recognizing emotions expressed by body pose: A biologically inspired neural model," *Neural Networks*, vol. 21, no. 9, pp. 1238–1246, 2008.

[4] M. Wöllmer, M. Kaiser, F. Eyben, B. Schuller, and G. Rigoll, "LSTM-Modeling of continuous emotions in an audiovisual affect recognition framework," *Image and Vision Computing*, vol. 31, no. 2, pp. 153–163, 2013.

[5] B. Schuller, S. Steidl, A. Batliner, J. Epps, F. Eyben, F. Ringeval, E. Marchi, , and Y. Zhang, "The INTERSPEECH 2014 Computational Paralinguistics Challenge: Cognitive & Physical Load," *Proc. Interspeech, ISCA*, pp. 427–431, Sep. 2014.

[6] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism," *Proc. Interspeech, ISCA*, pp. 148–152, Aug. 2013.

[7] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, "The INTERSPEECH 2011 Speaker State Challenge," *Proc. Interspeech, ISCA*, pp. 3201–3204, Aug. 2011.

[8] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, 2015.

[9] B. W. Schuller and F. Weninger, "Ten recent trends in computational paralinguistics," in *Cognitive Behavioural Systems*, ser. Lecture Notes in Computer Science, A. Esposito, A. M. Esposito, A. Vinciarelli, R. Hoffmann, and V. C. Müller, Eds. Springer Berlin Heidelberg, 2012, vol. 7403, pp. 35–49.

[10] H. A. Elfenbein and N. Ambady, "On the universality and cultural specificity of emotion recognition: A meta-analysis," *Psychological Bulletin*, vol. 128, no. 2, pp. 203–235, 2002.

[11] J. A. Russell, "Pancultural aspects of the human conceptual organization of emotions," *Journal of Personality and Social Psychology*, vol. 45, no. 6, pp. 1281–1288, 1983.

[12] P. Ekman, W. Friesen, and P. Ellsworth, Eds., *Emotion in the Human Face*. New York: Pergamon Press Inc., 1972.

[13] M. T. Riviello, A. Esposito, and K. Vicsi, "A Cross-Cultural Study on the Perception of Emotions: How Hungarian Subjects Evaluate American and Italian Emotional Expressions," in *Cognitive Behavioural Systems*, ser. Lecture Notes in Computer Science, A. Esposito, A. M. Esposito, A. Vinciarelli, R. Hoffmann, and V. C. Müller, Eds. Springer Berlin Heidelberg, 2012, vol. 7403, pp. 424–433.

[14] M. D. Pell, L. Monetta, S. Paulmann, and S. A. Kotz, "Recognizing Emotions in a Foreign Language," *Journal of Nonverbal Behavior*, vol. 33, no. 2, pp. 107–120, 2009.

[15] K. R. Scherer, R. Banse, and H. G. Wallbott, "Emotion Inferences from Vocal Expression Correlate Across Languages and Cultures," *Journal of Cross-Cultural Psychology*, vol. 32, no. 1, pp. 76–92, 2001.

[16] H. Argstatter, "Perception of basic emotions in music: Culture-specific or multicultural?" *Psychology of Music*, pp. 1–17, 2015.

[17] B. Schuller, B. Vlasenko, F. Eyben, M. Wollmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, "Cross-Corpus Acoustic Emotion Recognition: Variances and Strategies," *Affective Computing, IEEE Transactions on*, vol. 1, no. 2, pp. 119–131, 2010.

[18] M. Borchert and A. Dusterhoft, "Emotions in speech - experiments with prosody and quality features in speech for use in categorical and dimensional emotion recognition environments," *Proc. IEEE International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE)*, pp. 147–151, Oct. 2005.

[19] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing Emotions in Speech," *Proc. Fourth International Conference on Spoken Language Processing (ICSLP)*, vol. 3, pp. 1970–1973, Oct. 1996.

[20] A. Noguerias, A. Moreno, A. Bonafonte, and J. Mariño, "Speech Emotion Recognition Using Hidden Markov Models," *Proc. European Conference on Speech Communication and Technology (Eurospeech)*, pp. 2679–2682, Sep. 2001.

[21] 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 & Language*, vol. 25, no. 1, pp. 4–28, 2011.

[22] M. El Ayadi, M. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[23] H. Cao, A. Savran, R. Verma, and A. Nenkova, "Acoustic and lexical representations for affect prediction in spontaneous conversations," *Computer Speech & Language*, vol. 29, no. 1, pp. 203 – 217, 2015.

[24] M. El Ayadi, M. Kamel, and F. Karray, "Speech Emotion Recognition using Gaussian Mixture Vector Autoregressive Models," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, pp. IV–957–IV–960, Apr. 2007.

[25] Y.-L. Lin and G. Wei, "Speech emotion recognition based on HMM and SVM," *Proc. International Conference on Machine Learning and Cybernetics*, vol. 8, pp. 4898–4901, Aug. 2005.

[26] J. Wagner, T. Vogt, and E. André, "A Systematic Comparison of Different HMM Designs for Emotion Recognition from Acted and Spontaneous Speech," in *Affective Computing and Intelligent Interaction*, ser. Lecture Notes in Computer Science, A. Paiva, R. Prada, and R. Picard, Eds. Springer Berlin Heidelberg, 2007, vol. 4738, pp. 114–125.

[27] C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," *Proc. Interspeech, ISCA*, pp. 320–323, Sep. 2009.

[28] Z. Zeng, M. Pantic, G. Roisman, and T. Huang, "A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, Jan. 2009.

[29] E. M. Albornoz, D. H. Milone, and H. L. Rufiner, "Spoken emotion recognition using hierarchical classifiers," *Computer Speech & Language*, vol. 25, no. 3, pp. 556–570, 2011.

[30] D. Morrison, R. Wang, and L. C. De Silva, "Ensemble methods for spoken emotion recognition in call-centres," *Speech Communication*, vol. 49, no. 2, pp. 98–112, 2007.

[31] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," *Proc. Interspeech, ISCA*, pp. 312–315, Sep. 2009.

[32] A. Stuhlsatz, C. Meyer, F. Eyben, T. ZieIke, G. Meier, and B. Schuller, "Deep neural networks for acoustic emotion recognition: Raising the benchmarks," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5688–5691, May 2011.

[33] I. Lefter, L. J. Rothkrantz, P. Wiggers, and D. A. van Leeuwen, "Emotion recognition from speech by combining databases and fusion of classifiers," in *Text, Speech and Dialogue*, ser. Lecture Notes in Computer Science, P. Sojka, A. Horák, I. Kopecek, and K. Pala, Eds. Springer Berlin Heidelberg, 2010, vol. 6231, pp. 353–360.

[34] B. Schuller, Z. Zhang, F. Weninger, and G. Rigoll, "Using Multiple Databases for Training in Emotion Recognition: To Unite or to Vote?" *Proc. Interspeech, ISCA*, pp. 1553–1556, Aug. 2011.

[35] S. M. Feraru, D. Schuller, and B. Schuller, "Cross-language acoustic emotion recognition: An overview and some tendencies," *Proc. Affective Computing and Intelligent Interaction (ACII)*, pp. 125–131, Sep. 2015.

[36] F. Eyben, A. Batliner, B. Schuller, D. Seppi, and S. Steidl, "Cross-Corpus classification of realistic emotions – some pilot experiments," *Proc. Third International Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*, pp. 77–82, May 2010.

[37] B. Zhang, G. Essl, and E. M. Provost, "Recognizing emotion from singing and speaking using shared models," *Proc. Affective Computing and Intelligent Interaction (ACII)*, pp. 139–145, Sep. 2015.

[38] E. Mower, A. Metallinou, C.-C. Lee, A. Kazemzadeh, C. Busso, S. Lee, and S. Narayanan, "Interpreting ambiguous emotional expressions," *Proc. Third International Conference on Affective Computing and Intelligent Interaction and Workshops (ACII)*, pp. 1–8, Sep. 2009.

[39] S. Steidl, M. Levit, A. Batliner, E. Noth, and H. Niemann, ""Of All Things the Measure Is Man": Automatic Classification of Emotions and Inter-Labeler Consistency," *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, pp. 317–320, Mar. 2005.

[40] L. Devillers, L. Vidrascu, and L. Lamel, "Challenges in real-life emotion annotation and machine learning based detection," *Neural Networks*, vol. 18, no. 4, pp. 407–422, 2005.

[41] E. Mower, M. Matarić, and S. Narayanan, "A Framework for Automatic Human Emotion Classification Using Emotion Profiles," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1057–1070, Jul. 2011.

[42] Y. Attabi and P. Dumouchel, "Anchor models for emotion recognition from speech," *IEEE Transactions on Affective Computing*, vol. 4, no. 3, pp. 280–290, Jul. 2013.

[43] I. Lopez-Moreno, D. Ramos, J. Gonzalez-Rodriguez, and D. T. Toledano, "Anchor-model fusion for language recognition," *Proc. Interspeech, ISCA*, pp. 727–730, Sep. 2008.

[44] M. Collet, Y. Mami, D. Charlet, and F. Bimbot, "Probabilistic anchor models approach for speaker verification," *Proc. Interspeech, ISCA*, pp. 2005–2008, Sep. 2005.

[45] Y. H. Chin, S. H. Lin, C. H. Lin, E. Siahaan, A. Frisky, and J. C. Wang, "Emotion Profile-Based Music Recommendation," *Proc. 7th International Conference on Ubi-Media Computing and Workshops (UMEDIA)*, pp. 111–114, Jul. 2014.

[46] E. Mower and S. Narayanan, "A hierarchical static-dynamic framework for emotion classification," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2372–2375, May 2011.

[47] E. Mower, K. J. Han, S. Lee, and S. S. Narayanan, "A cluster-profile representation of emotion using agglomerative hierarchical clustering," *Proc. Interspeech, ISCA*, pp. 797–800, Sep. 2010.

[48] E. M. Provost, "Identifying salient sub-utterance emotion dynamics using flexible units and estimates of affective flow," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3682–3686, May 2013.

[49] Y. Kim and E. Provost, "Emotion classification via utterance-level dynamics: A pattern-based approach to characterizing affective expressions," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3677–3681, May 2013.

[50] C. Ortego-Resa, I. Lopez-Moreno, D. Ramos, and J. Gonzalez-Rodriguez, "Anchor Model Fusion for Emotion Recognition in Speech," in *Biometric ID Management and Multimodal Communication*, ser. Lecture Notes in Computer Science, J. Fiérrez-Aguilar, J. Ortega-Garcia, A. Esposito, A. Drygajlo, and M. Faúndez-Zanuy, Eds. Springer Berlin Heidelberg, 2009, vol. 5707, pp. 49–56.

[51] H. Cao, R. Verma, and A. Nenkova, "Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech," *Computer Speech & Language*, vol. 29, no. 1, pp. 186–202, 2015.

[52] Y. Kim and E. Mower Provost, "Leveraging inter-rater agreement for audio-visual emotion recognition," *Proc. Affective Computing and Intelligent Interaction (ACII)*, pp. 553–559, Sep. 2015.

[53] C. Zhang and Y. Ma, *Ensemble Machine Learning: Methods and Applications*. Springer US, 2012.

[54] P. Ekman, E. R. Sorenson, and W. V. Friesen, "Pan-cultural elements in facial displays of emotions," *Science*, vol. 164, no. 3875, pp. 86–88, Apr. 1969.

[55] C.-H. Wu, J.-C. Lin, and W.-L. Wei, "Survey on audiovisual emotion recognition: databases, features, and data fusion strategies," *APSIPA Transactions on Signal and Information Processing*, vol. 3, p. e12, 2014.

[56] Y. Wang and L. Guan, "Recognizing Human Emotional State From Audiovisual Signals," *IEEE Transactions on Multimedia*, vol. 10, no. 5, pp. 936–946, 2008.

[57] Y. Wang, L. Guan, and A. N. Venetsanopoulos, "Kernel Cross-Modal Factor Analysis for Information Fusion With Application to Bimodal Emotion Recognition," *IEEE Transactions on Multimedia*, vol. 14, no. 3, pp. 597–607, 2012.

[58] Y. Tie and L. Guan, "A Deformable 3-D Facial Expression Model for Dynamic Human Emotional State Recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 1, pp. 142–157, 2013.

[59] C. S. Ooi, K. P. Seng, L.-M. Ang, and L. W. Chew, "A new approach of audio emotion recognition," *Expert Systems with Applications*, vol. 41, no. 13, pp. 5858–5869, 2014.

[60] C. Plapous, C. Marro, and P. Scalart, "Improved Signal-to-Noise Ratio Estimation for Speech Enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 2098–2108, 2006.

[61] T. Giannakopoulos and A. Pikrakis, *Introduction to Audio Analysis: A MATLAB® Approach*, 1st ed. Oxford: Academic Press, 2014.

[62] B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture," May 2004, pp. I–577–580.

[63] J. Adell Mercado, A. Bonafonte Cávez, and D. Escudero Mancebo, "Analysis of prosodic features: towards modelling of emotional and pragmatic attributes of speech," *Procesamiento de Lenguaje Natural*, no. 35, pp. 277–283, 2005.

[64] X. Yang, K. Wang, and S. A. Shamma, "Auditory representations of acoustic signals," *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 824–839, 1992.

[65] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 887–906, 2005.

[66] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001, software available at www.csie.ntu.edu.tw/~cjlin/libsvm/.

[67] Y.-W. Chen and C.-J. Lin, "Combining svms with various feature selection strategies," in *Feature extraction*. Springer, 2006, pp. 315–324.

Enrique M. Albornoz received the Engineering informatics degree (Hons.) from National University of Litoral (UNL), Argentina, in 2006, and the Ph.D. degree on Engineering oriented to Computational Intelligence, Signals and Systems from National University of Litoral (UNL), Argentina, in 2011. He is with the Research Institute for Signals, Systems and Computational Intelligence - sinc(i) (UNL-CONICET) since 2004. In 2007 he started as Professor in the Department of Informatics at National University of Litoral (UNL) and since 2014 he is a Research Scientist at the National Scientific and Technical Research Council (CONICET). His research interests include statistical learning, pattern recognition, signal and image processing, with applications to speech recognition, affective computing and biomedical signals.

Diego H. Milone received the Bioengineering degree (Hons.) from National University of Entre Ríos (UNER), Argentina, in 1998, and the Ph.D. degree in Microelectronics and Computer Architectures from Granada University, Spain, in 2003. He was with the Department of Bioengineering and the Department of Mathematics and Informatics at UNER from 1995 to 2002. Since 2003 he is Full Professor in the Department of Informatics at National University of Litoral (UNL). From 2009 to 2011 was Director of the Department of Informatics and from 2010 to 2014 was Assistant Dean for Science and Technology. He is Director of the Research Institute for Signals, Systems and Computational Intelligence, sinc(i), FICH-UNL/CONICET. Since 2006 he is a Research Scientist at the National Scientific and Technical Research Council (CONICET). His research interests include statistical learning, pattern recognition, signal processing, neural and evolutionary computing, with applications to speech recognition, affective computing, biomedical signals and bioinformatics.