Biomedical Signal Processing and Control 7 (2012) 79–87

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Acoustic feature selection and classification of emotions in speech using a 3D continuous emotion model

Humberto Pérez-Espinosa*, Carlos A. Reyes-García, Luis Villaseñor-Pineda

Instituto Nacional de Astrofísica, Óptica y Electrónica, Luis Erique Erro 1, Tonantzintla, Puebla 72840, Mexico

ARTICLE INFO

Article history:
Received 15 September 2010
Received in revised form 23 February 2011
Accepted 24 February 2011
Available online 3 April 2011

Keywords:
Automatic emotion recognition
Continuous emotion model
Feature selection

ABSTRACT

In this paper we report the results obtained from experiments with a database of emotional speech in English in order to find the most important acoustic features to estimate Emotion Primitives which determine the emotional content on speech. We are interested in exploiting the potential benefits of continuous emotion models, so in this paper we demonstrate the feasibility of applying this approach to annotation of emotional speech and we explore ways to take advantage of this kind of annotation to improve the automatic classification of basic emotions.

1. Introduction

Emotions are very important in our everyday life, they are present in everything we do. There is a continuous interaction between emotions, behavior and thoughts, in such a way that they constantly influence each other. Emotions are a great source of information in communication and interaction among people, they are assimilated intuitively.

The applications of emotion recognition encompass many fields, for example, as a supporting technology in medical areas such as psychology, neurology and caring of aged and impaired people. Automatic emotion recognition based on biomedical signals, facial and vocal expressions has been applied to diagnosis and following-up of progressive neurological disorders, specifically Huntington's and Parkinson's diseases [1]. These pathologies are characterized by a deficit in emotional processing of fear and disgust and, thus, the system could be utilized to determine the subject reaction/or absence of reaction to specific emotions, helping the health professionals to gain a better understanding of these disorders. Furthermore, the system could be used as a reference to evaluate the patients' response to certain medicines. Another important medical application is remote medical support. These kinds of environments enable communication between medical professionals and patients for cases of regular monitoring and emergency situations. In this scenario the system recognizes patient's emotional states and then transmits data indicating the patient is experiencing depression or sadness, health-care providers monitoring them will be better prepared to respond. Such a system has the potential to improve patient satisfaction and health [2–4]. It can also be an asset for disabled people who have difficulties with communication. Hearing-impaired people who are not profoundly deaf can use residual hearing to communicate with other people and learn to speak with emotions making communication more complete and understandable. In these cases, emotion recognition engines can be used as an important element of a computer-assisted emotional speech training system [5]. For hearing-impaired people, it could provide an easier way to learn how to speak with emotion more naturally or help speech therapist to guide them to express correctly emotions in speech. Emotion recognition arouses great interest in the interface design given that recognizing and understanding emotions automatically is one of the key steps towards emotional intelligence in Human–Computer Interaction (HCI). The need for automatic emotion recognition has emerged due to the tendency towards a more natural interaction between humans and computers. Affective computing is a topic within HCI that encompass these research tendency trying to endow computers with the ability to detect, recognize, model and take into account user's emotional state that plays a role of paramount importance in the way humans make decisions [6]. Emotions are essential for human thought processes that influence interactions between people and intelligent systems.

In the area of automatic emotion recognition mainly two annotation schemes have been used to capture and describe the emotional content in speech: discrete and continuous approaches. Discrete approach is based on the concept of basic emotions such as anger, joy, and sadness, that are the most intense form of emo-

* Corresponding author.
E-mail addresses: humbertop@inaoep.mx (H. Pérez-Espinosa), kargaxxi@inaoep.mx (C.A. Reyes-García), villasen@inaoep.mx (L. Villaseñor-Pineda).

1746-8094/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bspc.2011.02.008

tions from which all other emotions are generated by variations or combinations of them. They assume the existence of universal emotions that can be clearly distinguished from one another by most people. On the other hand, continuous approach represent emotional states using a continuous multidimensional space. Emotions are represented by regions in an *n*-dimensional space where each dimension represents an *Emotion Primitive*. Emotion Primitives are subjective properties shown by all emotions. The most widely accepted primitives are Arousal and Valence. Valence describes how negative or positive is a specific emotion. Arousal, also called Activation, describes the internal excitement of an individual and ranges from being very calm to be very active. Also, three-dimensional models have been proposed. The most common three-dimensional model includes Valence, Activation as well as Dominance [7]. This additional primitive describes the degree of control that the individual intends to take on the situation, or in other words, how strong or weak the individual seems to be.

Both approaches, discrete and continuous, provide complementary information about the emotional manifestations observed in individuals [8]. Discrete categorization allows a more particularized representation of emotions in applications where it is needed to recognize a predefined set of emotions. However, this approach ignores most of the spectrum of human emotional expressions. In contrast, continuous models allow the representation of any emotional state and subtle emotions, but it has been found that it is difficult to estimate with high precision the Emotion Primitives based only on acoustic information. Some authors have begun to research about how to take advantage of this theory [9–11] to estimate more adequately the emotional content in speech. Both approaches are closely related, by assessing the emotional content in speech using one of these two schemes we can infer its counterparts in the other scheme. For instance, if an utterance is evaluated as anger we may infer that the utterance would have a low value for Valence and high for Activation and Dominance. Conversely, if an utterance is evaluated with low Valence and high Activation and Dominance we could infer that this is Anger. Several authors [12–14] have worked on the analysis of the most important acoustic features from the point of view of discrete categorization; however, they have not yet studied with the same depth the importance of acoustic attributes from the continuous models point of view. We believe that the continuous approach has great potential to model the occurrence of emotions in the real world. This three-dimensional continuous model is adopted in this paper. As a first step towards exploiting the continuous approach we analyze the most important acoustic features to automatically estimate Emotion Primitives in speech. Then, we will be able to use this estimation in order to locate the individual's emotional state in the multidimensional space, and if necessary, to map it to a basic emotion. In this work we apply these ideas to improve the automatic emotion recognition from acoustic information. We perform some experiments in order to find the most important acoustic features for estimating Emotion Primitives in speech and then we propose a method that uses these estimations to determine the individual's emotional state mapping it to a basic emotion. The remainder of this paper is organized as follows. First, we describe the database in Section 2. Next, in Section 3 we describe the acoustic features and how we are extracting them from speech signals. The filters we applied to select the best instances are explained in Section 4. In Section 5 we propose and describe two ways of finding the best acoustic features for Emotion Primitives Estimation. Experimental results and discussions about feature selection are provided in Section 6. In Section 7 we propose a way of applying the automatic Emotion Primitive estimation in order to classify basic emotions. In Section 8, we present and discuss the results obtained from basic emotions classification. Finally, the conclusion of this study and future work are discussed in Section 9.

2. Database

For the purposes of this work it is necessary to have a database labeled with Emotion Primitives, namely Valence, Activation and Dominance. In addition, we need basic emotions annotations for each sample to validate the Emotion Primitives estimation accuracy by evaluating the mapping done from continuous to discrete approach. There are many databases labeled with emotional categories such as FAU Aibo [15], Berlin Database of Emotional Speech [16] Spanish Emotional Speech [17] and a few that are labeled with Emotion Primitives VAM Corpus [18], IEMOCAP Database [8]. The database we used is labeled with common discrete categories and Emotion Primitives. This database is called IEMOCAP [8] (Interactive Emotional Dyadic Motion Capture Database). It was collected at the Speech Analysis and Interpretation Laboratory at the University of Southern California and it is in English. It was recorded from ten actors in male–woman pairs. IEMOCAP includes information about head, face and hands motion as well as audio and video offering detailed information about facial expression and gestures. To generate an emotional dialogue they designed two scenarios. In the first one, the actors were following a script, while in the second one, the actors improvised according to a preset situation. The complete corpus contains about 12 h, although at the moment only the first session, i.e., the interaction of a pair of actors has been released. The utterances were labeled with the emotional categories: happiness, anger, sadness, frustration, fear, surprise and neutral. The categories "other" and not "identified" were also included. To evaluate Emotion Primitives an integer value between one and five was annotated for each primitive, Valence (1 – negative, 5 – positive), Activation (1 – calm, 5 – excited), and Dominance (1 – weak, 5 – strong). The characteristics of this database make it very interesting for the purposes of our work since the annotation includes the two most important approaches. Attention has been given to spontaneous interaction, in spite of using actors, in ideal conditions for recording. The database shows a significant diversity of emotional states. We use only the segmented audio in turns of the first session for the categories: anger, happiness, neutral and surprise. In total, there are 1.820 instances.

3. Features extraction

We extracted acoustic features from the speech signal using two programs for acoustic analysis Praat [19] and OpenEar [20]. We evaluated two sets of features; one of them was obtained through a selective approach, i.e., based on a study taking into account the features that could be useful, the features that have been successful in related works and features used for other similar tasks. The second feature set was obtained by applying a brute force approach, i.e., generating a large amount of them hoping that some will be found useful. The Selective Feature Set is a set of features that we have been building over our research [21,22] and was extracted with the software Praat. We designed this set of features representing several voice aspects, we included the traditional attributes associated with prosody, e.g., duration, pitch and energy. Others which have shown good results in tasks like speech recognition, speaker recognition, classification of baby cry [23], language recognition [24], and pathology detection in voice [25,26] and that we believe can be successful in emotion recognition. Table 1 shows the number of acoustic features that were included in each feature group. We divide the three types of features we include in: prosodic, spectral and voice quality. Prosodic features describe suprasegmental phenomena, i.e., speech units larger than phonemes, such as pitch, loudness, speed, duration, pauses and rhythm. Prosody is a rich source of information in speech processing because it has a very important paralinguistic information that complements the mes-

Table 1
Selective approach feature set.

Group	Feature type	Number of features
Prosodic		
Times	Elocution times	8
F0	Melodic contour	9
Energy	Energy contour	12
Voice quality		
Voice quality	Quality descriptors	24
Voice quality	Articulation	12
Spectral		
LPC	Fast Fourier transform	4
LPC	Long term average	5
LPC	Wavelets	6
MFCC	MFCC	96
Cochleagram	Cochleagram	96
LPC	LPC	96
Total		368

Table 2
Brute force approach feature set.

Group	Feature type	Number of features
Prosodic		
Energy	LOG energy	117
Times	Zero crossing rate	117
PoV	Probability of voicing	117
F0	F0	234
Spectral		
MFCC	MFCC	1521
MEL	MEL spectrum	3042
SEB	Spectral energy in bands	469
SROP	Spectral roll off point	468
SFlux	Spectral flux	117
SC	Spectral centroid	117
SpecMaxMin	Spectral max and min	233
Total		6552

sage with an intention that can reflect an attitude or emotional state [27]. These type of features are the most commonly used in speech emotion recognition. We subdivide prosodic features in elocution times, melodic contour and energy contour. The second type of features is voice quality that gives the primary distinction to a given speaker's voice when prosodic aspects are excluded. Some of the descriptions of voice quality are harshness, breathiness and nasality. Some authors [15] has studied the importance of voice quality stating that the key to the vocal differentiation of discrete emotions seems to be voice quality. We included the two most popular voice quality descriptors that are jitter and shimmer. We also included other voice quality descriptors that have been related to the GRBAS scale in related work [25–29]. Some of these features have never been used in speech emotion recognition. For example, the energy differences between frequency bands and the ratio between frequency bands were used by [25] to discriminate between pathological and normal voices. Increases and decreases in energy peaks were used by [26] for automatic detection of vocal fry. Articulation is also an important parameter to measure voice quality. We included some statistical measures of the first four formants as articulatory descriptors. Spectral features describe the characteristics of a speech signal in the frequency domain besides F0 like harmonics and formants [15]. We included several types of spectral representations. Some of these representations have never been used in emotion recognition as cochleagrams that have been used for infant cry classification [23] and others that have studied very little for this task as Wavelets [30]. A cochleagram [19] represents the excitation of the auditory nerve filaments of the basilar membrane, which is situated in the cochlea in the inner ear. This excitation is represented as a function over time (s) and Bark frequency that is a psychoacoustic. A cochleagram also models the sound volume and frequency masking. Frequency masking in the ear, happens when we hear two sounds of different intensity at the same time, the weakest sound is not distinguished as the brain only processes the masking sound. Features based on cochleagrams have been used for speech recognition with good results. In the work of Shamma et al. [31] cochleagrams were used for speech recognition at phoneme level, surpassing the results obtained by LPC features. Wavelets are an alternative to the Fourier transform. Wavelet transform allows a good resolution at low frequencies. We also included features widely used in speech processing as MFCCs (Mel-frequency cepstral coefficients) [32] that represents speech perception based on human hearing and have been successfully used to discriminate phonemes. MFCCs have shown that they not only are useful to determine what is said but, also how it is said.

The brute force feature set was extracted using the software OpenEar. We extract a total of 6552 features including first-order functionals of low-level descriptors (LLD) such as FFT-Spectrum, Mel-Spectrum, MFCC, Pitch (Fundamental Frequency F0 via ACF), Energy, Spectral, LSP.39 functionals such as Extremes, Regression, Moments, Percentiles, Crossings, Peaks, Means were applied. Table 2 shows features extracted by brute force approach. In emotion recognition from speech there are two approaches for feature processing static and dynamic approach. The dynamic processing captures information about how features evolve over time. By the other side, static processing avoids overfitting in the phonetic modeling by applying statistical functions on low level descriptors in periods of time. Static processing is more common on emotion recognition from speech. However, dynamic processing has showed good results in recent years [33–35]. And even new methods have been proposed to represent the signal properties associated with the relationship between expressiveness and voice quality [36]. Our spectral features include static coefficients that describe the spectral properties within one frame where the signal is approximately stationary. Our feature set includes dynamic features, which describe the behavior of the static features over time. For this purpose, the first and the second derivative of the static features in the brute force set were calculated.

4. Instance selection

Doing an inspection on the database instances, we realized that there were problematic instances; we thought that our machine learning algorithm would have a better performance by selecting the most appropriate and congruent instances representing the properties of continuous and discrete approaches. Our initial data set consists of 1820 instances. We worked only with the more represented classes, so we first selected the instances from the four classes with more examples. We choose the instances from these four classes to enable the comparison of our results to the work done by [37], where they use only these four classes. After applying this filter we are left with 942 instances, 20% was reserved for final testing. The experiments reported in Sections 6, 7 and 8 were done using the 80% equivalent to 753 instances. Another filter was applied to these instances which consisted of removing all instances that have the same annotation for each of the three primitives, but different annotation for basic emotion, to be regarded as contradictory instances that add noise to our learning process. For example, if annotations for an instance are (Valence = 2, Activation = 4, Dominance = 3, Emotion = Angry) and the annotations for another instance are (Valence = 2, Activation = 4, Dominance = 3, Emotion = Happiness) every instance annotated with Valence = 2, Activation = 4 and Dominance = 3 is eliminated from our data set. After this filtering, we were working with a set of 401 instances in the feature selection process.

Table 3
Correlation index obtained for primitives estimation using the whole feature set.

Emotion Primitive	Correlation index
Valence	0.0151
Activation	0.0095
Dominance	−0.001

5. Feature selection

The need of finding the best feature subsets for building our learning models arises given the low correlation obtained in the estimation of primitives using a trained model for the full set of 6920 acoustic features. Correlation coefficient measures the quality of the estimated variable determining the strength and the direction of a linear relationship between the estimated and the actual value of the variable. The closer the coefficient is to either −1 or 1, the stronger the correlation between the variables. As it approaches zero there is less of a relationship. Table 3 shows the correlation coefficient obtained when estimating the value of the Emotion Primitives with a model built from 942 instances (using only the four classes with more instances) and 6920 attributes. As we can see, correlation is very low and therefore, the learnt models from these data are not useful. Having too many features in relation to the number of instances complicates the classifiers task, SVM in this case, impeding a proper prediction model. Although many attributes can enhance discrimination power, in practice, with a limited amount of data, an excessive amount of attributes significantly delays the learning process and often results in over-fitting.

Initially, we tested different attribute selectors such as SubSetEval [38] that evaluates the value of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them, preferring the subsets of features that are highly correlated with a class and lowly correlated with each other. We also tested ReliefAttribute [38] The key idea of ReliefAttribute is to estimate attributes according to the values that distinguish the closer instances. For a given instance, ReliefAttribute seeks two nearest neighbors: one from the same class and another from a different class. Good attributes must be on one hand, different values between instances of different classes and on the other hand, the same values for instances of the same class. Using these techniques we could not improve the correlation results for primitives estimation. Due to these problems, it is necessary to devise a way to select the best features among a large number of them, bearing in mind that we have a few instances. We propose two schemes for selecting attributes working on smaller feature sets with the idea of avoiding the search for the best features on the whole set.

5.1. Scheme 1: ungrouped feature selection

Fig. 1 shows the processes applied in this scheme. We started from an initial feature set obtained from a feature selection process applied to the VAM (*Vera am Mittag*) German spontaneous speech database [18]. This initial feature set achieved good correlation results in the estimation of Emotional Primitives in VAM

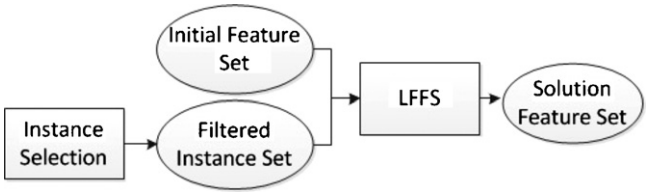

Fig. 1. Ungrouped feature selection scheme.

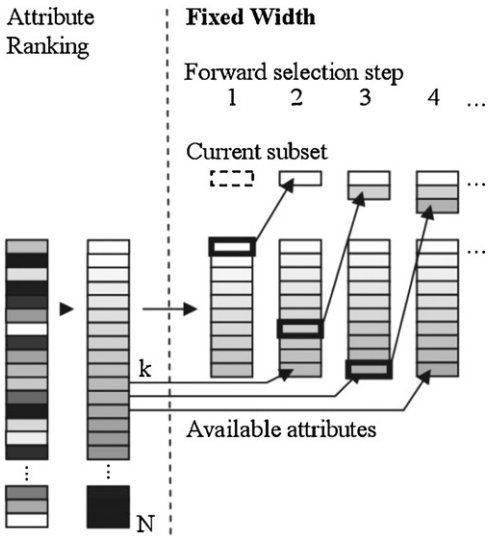

Fig. 2. Linear forward selection with fixed width (taken from [40]).

database. This selection process was carried out with 252 features obtained by selective approach and 949 instances [22]. Later, we conducted the instance selection process explained in Section 4. Finally, we applied the feature selection process known as Linear Floating Forward Selection (LFFS) [39] which makes a Hill-Climbing search, starting with the empty set or with a predefined set, evaluates all possible inclusions of a single attribute to the solution subset. At each step the attribute with the best evaluation is added. The search ends when there are no inclusions that improve the evaluation. In addition, LFFS dynamically changes the number of features included or eliminated at each step. In our experiments we use the LFFS modality called fixed width shown in Fig. 2. In this mode the search is not performed on the whole feature set. Initially, the *k* best features are selected and the rest are removed. In each iteration the features added to the solution set are replaced by features taken from those that had been eliminated. This scheme processes are repeated for each Emotion Primitive.

5.2. Scheme 2: grouped feature selection

In this scheme, the idea is to divide the whole feature set into smaller groups according to the acoustic properties they represent. Feature groups are shown in Tables 1 and 2. Fig. 3 shows the steps followed in this scheme. First, we applied the instance selection filter explained in Section 4. Second, we divided the whole data set into smaller sets grouping features sharing the same acoustic properties (see Tables 1 and 2) and we applied LFFS, unlike Scheme 1, this time the search process by LFFS started from the empty set. Third, the selected features for each group were put together in a final set. The three steps in this selection scheme by groups of features are repeated for each Emotion Primitive.

6. Feature selection results

All the results of the learning experiments in this paper were obtained using Support Vector Machines (SVM) and validated by 10-Fold Cross-Validation.

The metrics used to measure the importance of feature groups are: correlation coefficient, share and portion. The correlation coefficient is the most common parameter to measure the machine learning algorithms performance on regression tasks, as in our case. We use share and portion that are measures proposed in [14] to assess the impact of different types of features on the performance of automatic recognition of discrete categorical emotions.

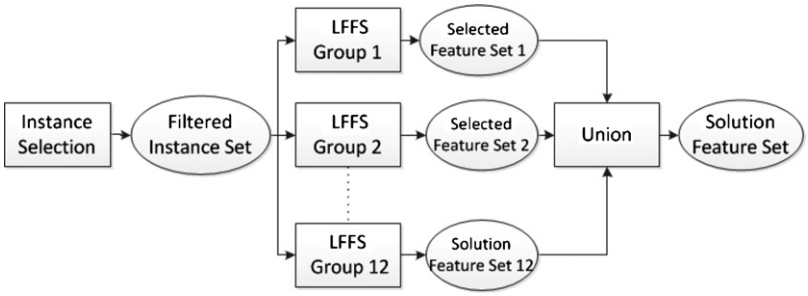

Fig. 3. Grouped feature selection scheme.

Correlation coefficient: Indicates the strength and direction of the linear relationship between the annotated primitives and the estimated primitives by the trained model. It is our main metric to measure the classification results.

Share: Shows the contribution of types of features. For example, if 28 features are selected from the group times and 150 were selected in total then.

$$\text{Share} = \frac{(28 \times 100)}{150} = 18.7$$

Portion: Shows the contribution of types of features weighted by the number of features per type. For example, if 28 features are selected from a total of 125 features of the times set then:

$$\text{Portion} = \frac{(28 \times 100)}{125} = 22.4$$

Having identified the best acoustic feature sets we constructed individual classifiers to estimate each Emotion Primitive. Table 4 shows the evaluation results of the instances/attributes selection scheme illustrated in Fig. 1. The second column shows the results when using all attributes and all instances in the learning process. As we can see the correlation coefficient is too low. The third column shows the results when the learning process is performed using the best features for each primitive proposed by [22]. The fourth column shows the results when the learning process uses the same features as the previous experiment, but we filtered instances as described in Section 4. Finally, the fifth column shows the results after filtering instances and applying the feature selection method LFFS. As we can see the improvement in results was gradual after applying each data processing.

The two feature selection schemes were applied to each of the three Emotion Primitives, that is, we ran six times the feature selection process, obtaining six different feature sets. This paper shows the results of the best subsets for each primitive. Tables 5–7 reflect the effectiveness of each feature set and the differences among groups. The groups PoV and SC are not shown in these Tables because no features were selected from them. It is important to note that the correlation, share and portion shown in these tables are obtained by decomposing in groups of features the solution set found by the selection Schemes 1 and 2 and evaluating them separately with these metrics. The higest value for each metric is marked in bold. While feature selection Scheme 2 ensures that at least one attribute of each group will be included, Scheme 1 may not include any element of certain groups in the solution set. The last row shows the total number of features selected for the Emotion Primitive. In the experiments for Valence with the selection Scheme

Table 4
Results for each step of the ungrouped feature selection scheme. Each column shows the number of features/correlation coefficient.

Emotion Primitive	Baseline results		Scheme 1 results	
	All attributes	Initial feature selection	Instance selection	LFFS
Valence	6920/0.0151	56/0.5188	56/0.5597	62/0.6189
Activation	6920/0.0095	67/0.7463	67/0.7861	60/0.7969
Dominance	6920/−0.001	23/0.6536	23/0.7117	31/0.7437

Table 5
Scheme 1/scheme 2 feature selection results – Valence.

Feature group	Total	Selected	Correlation	Share	Portion
Voice quality	36	6/5	0.29/0.36	9.67/7.14	**16.66**/13.88
Times	125	1/5	0.33/0.36	1.61/7.14	0.80/4.00
Cochleagrams	96	8/10	0.40/0.51	12.90/14.28	8.33/10.41
LPC	111	13/8	0.34/0.53	20.96/11.42	11.71/7.20
SFlux	117	0/5	−/0.48	0/7.14	0/4.27
Energy	129	0/6	−/0.37	0/8.57	0/4.65
F0	243	0/1	−/0.31	0/1.42	0/0.41
SpecMaxMin	234	0/1	−/0.18	0/1.42	0/0.42
SEB	234	0/5	−/0.45	0/7.14	0/2.13
SROP	468	0/6	−/0.35	0/8.57	0/1.28
MFCC	1617	27/13	0.48/0.49	**43.54**/18.57	1.67/0.80
MEL	3042	7/5	0.51/0.51	11.29/7.14	0.24/0.16
All	6920	62/70	0.61/**0.62**	100/100	0.89/1.04

Table 6
Scheme 1/scheme 2 feature selection results – Activation.

Feature group	Total	Selected	Correlation	Share	Portion
Voice quality	36	3/10	0.08/0.72	5.00/9.80	8.33/**27.77**
Times	125	1/10	0.16/0.71	1.66/9.80	0.80/8.00
Cochleagrams	96	24/10	0.77/0.78	**40.00**/9.80	25.00/10.41
LPC	111	4/12	0.61/0.78	6.66/11.76	3.60/10.81
SFlux	117	0/4	−/0.65	0.00/3.92	0.00/3.41
Energy	129	4/11	0.76/0.78	6.66/10.78	3.10/8.52
F0	243	1/9	0.29/0.65	1.667/8.82	0.412/3.70
SpecMaxMin	234	0/11	−/0.78	0.00/10.78	0.00/4.70
SEB	234	0/4	−/0.53	0.00/3.92	0.00/1.70
SROP	468	0/6	−/0.65	0.00/5.88	0.00/1.28
MFCC	1617	23/6	0.77/0.78	38.33/5.88	1.42/0.37
MEL	3042	0/9	−/0.73	0/8.82	0/0.29
All	6920	60/102	**0.79**/0.78	100/100	0.86/1.47

Table 7
Scheme 1/scheme 2 feature selection results – Dominance.

Feature group	Total	Selected	Correlation	Share	Portion
Voice quality	36	0/10	−/0.65	0.00/13.15	0.00/**27.77**
Times	125	2/6	0.35/61	6.45/7.89	1.60/4.80
Cochleagrams	96	7/8	0.70/72	22.58/10.52	7.29/8.33
LPC	111	1/7	0.66/72	3.22/9.21	0.90/6.30
SFlux	117	1/5	0.61/0.66	3.22/6.57	0.85/4.27
Energy	129	2/12	0.15/0.71	6.45/15.78	1.15/9.30
F0	243	3/4	0.22/0.59	9.67/5.26	1.23/1.64
SpecMaxMin	234	0/6	−/0.42	0.00/7.89	0.00/2.56
SEB	234	0/2	−/0.59	0.00/2.63	0.00/0.85
SROP	468	0/4	−/0.53	0.00/5.26	0.00/0.85
MFCC	1617	11/8	0.71	**35.48**/10.52	0.68/0.49
MEL	3042	4/4	0.68/0.69	12.90/5.26	0.13/0.13
All	6920	31/76	**0.74**/0.72	100/100	0.44/1.09

1, the group with higher correlation was MEL (0.5167), contributing to the solution set with 7 out of 62 features. The MEL portion is very low (0.230) there is a total of 3042 features belonging to this group. Share may be considered medium (11.290). This indicates that few MEL coefficients provide very important information. Another important group was MFCC with a correlation of 0.4877, indicating that the spectral information groups are important to estimate Valence. In the experiment for Valence with the selection Scheme 2, the best groups were LPC (0.5349) cochleagrams (0.5152) and MEL (0.5129). LPC and cochleagrams show a share (11.429–14.284) and portion (7.207–10.417) similar, meanwhile MEL shows a share (7.143) and portion (0.164) lower in comparison with the two mentioned groups. As in Scheme 1, MEL provides few features, but very important features. We realized that for Valence spectral type groups were the best for Schemes 1 and 2. No group by itself reached the correlation obtained with all groups together. We can see that the 8 selected attributes in Scheme 2 from LPC were much better (0.535) than the 13 attributes selected in Scheme 1 for the same group (0.342). The best correlation for Valence was obtained using the Scheme 2 reaching 0.6232.

For Activation the group MFCC has the best correlation in both selection schemes MFCC obtained 0.7795 in Scheme 1 and 0.7897 in Scheme 2. We can see that its share (38.333 and 5.882) and portion (1.422 and 0.371) are very different. We can see that using only the group MFCC we can reach similar results to those obtained using all groups (0.7964 and 0.787). These results clearly indicate the importance of this group to estimate Activation. As expected, the signal energy was very important for this primitive with a correlation of 0.7851 in Scheme 2 and 0.7639 in Scheme 1 as it is expected that people experiencing increased activity or excitement show a higher volume in its voice. Other important groups for Activation are cochleagrams and LPC. An interesting effect is that the group SpecMaxMin was important in the Scheme 2 with a correlation of 0.7831, with a high share, but in Scheme 1, this was not selected any feature from this group. It was expected that the group times was important for Activation since it has been found that the faster speech is perceived the more excited the individual is and the slower speech is perceived the more relaxed the individual is perceived to be [27].

Only one times feature was selected in the Scheme 1; its correlation was very low (0.167), whereas in Scheme 2, 10 features were selected obtaining a relatively high correlation (0.7128). In Scheme 2 the best results were not obtained using all selected attributes (0.787) but using only the ones selected by the brute force approach (0.7952). The best result for Activation was obtained using all groups in Scheme 1 0.7964. We infer that Activation major groups are spectral, MFCC, cochleagrams and energy.

In the experiments carried out with Dominance in Scheme 1 the best groups were MFCC (0.72) and cochleagrams (0.702) and in Scheme 2 LPC (0.7266) and cochleagrams (0.7244). In Scheme 2 the best results were obtained using only the features selected from group LPC (0.7266), surpassing the results obtained using all groups (0.7157). In Scheme 2 the best results were not obtained using all the features selected (0.7157) but only those selected by the selective approach (0.726). The best result for Dominance was obtained with Scheme 1 using all groups (0.7437). We can infer that in the case of Dominance the most important groups are spectral, MFCC and cochleagrams.

6.1. Selective vs. brute force

The first studies on automatic emotion recognition often opted for a selective feature extraction approach based on expert knowledge, usually with a small number of features. Today, with the emergence of tools that allow us to extract a large number of features and the availability of more computing power it is easier to apply a brute force approach. One of the points to consider in this paper is to compare the selective and brute force feature extraction approaches and to analyze which one could be better. Furthermore, we are interested in validating the work that we have done before following a selective approach [21,22].

Table 8
Selective vs. brute force feature extraction approach – scheme 1/scheme 2.

Approach	Total	Selected	Correlation	Share	Portion
Valence					
Selective	368	46/27	0.48/0.56	74.19/38.57	12.50/7.34
Brute force	6552	16/43	0.57/0.55	25.81/61.43	0.24/0.68
Activation					
Selective	368	56/37	0.79/0.79	93.33/36.28	15.22/10.05
Brute force	6552	4/65	0.76/0.80	6.67/63.72	0.06/1.03
Dominance					
Selective	368	13/29	0.72/0.73	41.93/38.16	3.53/7.88
Brute force	6552	18/47	0.73/0.70	58.06/61.84	0.28/0.74

Table 8 shows a comparison between feature extraction approaches (selective vs. brute force) as well as a comparison between the feature selection schemes here proposed. In all experiments, the selective and brute force extraction approaches got correlation coefficients very similar with the exception of the experiment done with Scheme 1 for Valence where the results with brute force (0.5715) were much better than with selective (0.4799). However, the selective approach share (74.194) was much higher than the brute force one (25.806) and the correlation using both groups was higher (0.6189) than the obtained for each group separately. It is very difficult to say which feature extraction scheme is better since they obtained similar results using the selected features from these groups separately and generally yield better results joining them into one set.

7. Basic emotion classification based on Emotion Primitives

Having identified the most relevant acoustic features for estimating the Emotion Primitives, the next step is to devise a way to use these estimations to discriminate emotional states in people. In this section, we want to strengthen the findings in the features study by demonstrating that, the continuous approach can actually help us to improve the automatic emotion classification from acoustic features based on a continuous emotion model. Section 2 describes the database we are working on. The corpus IEMOCAP was annotated with basic emotions and Emotion Primitives.

The classification scheme used to perform the experiments reported in this paper is illustrated in Fig. 4.

1. Acoustic features are extracted from the speech samples, as we have mentioned, we tested two sets of features, the selective set and the brute force set.
2. We applied a procedure called transformation, which consist in replacing the basic emotions labels with values corresponding to linguistic labels, for example, if an instance is labeled as anger that label is replaced by three labels, one for each primitive:
   - Low for Valence
   - High for Activation
   - High for Dominance

   This labeling was done according to Table 9. Since this database has a manual annotation of Emotion Primitives those annotations were used to build the table. We obtained the means of the annotated values per class from the annotations of the instances filtered according to the process described in Section 4.
3. After extracting the acoustic features and performing the transformation step, we had a data set consisting of acoustic feature vectors whose values to predict are the Emotion Primitives Valence, Activation and Dominance, the possible values that

Table 9
Primitive values calculated according to annotations made by IEMOCAP corpus evaluators. The values in italic are considered low and values in bold are considered high.

Basic emotion	Valence	Activation	Dominance
Anger	*2*	**4**	**3.5**
Happiness	**4.42**	**3.35**	*2.58*
Neutral	**2.75**	*3*	*2.5*
Sadness	*2*	*2*	*1.75*

these primitives can take are high or low. With this information, three models are trained to estimate the values of each Emotion Primitive. To build these models we used Support Vector Machines.

4. To perform the classification of basic emotions we extracted acoustic features, and apply the three models obtained in the previous step. In this way you get a high or low value for each of the instances. With these three values we constructed vectors whose attributes are the three primitives and the value to predict is the basic emotion of the original instance.
5. Finally we applied a machine learning algorithm (SVM) to construct a model to assign an emotion to each instance.

8. Basic emotions classification results

Table 10 shows the results when classifying Emotion Primitives according to high and low classes as described in Section 7 point 3 of the classification process description.

Table 11 shows a comparison of the results obtained by applying the process described in Fig. 4, using a discrete approach for the classification of emotions and the baseline results for the corpus with which we are working reported in [37]. We can see that it has achieved a best recall (number of correctly classified divided by number of total items) using our method on four emotions, anger, happiness and neutral, whereas sadness has a lower performance. These results show that by estimating Emotion Primitives we can discriminate with an adequate accuracy prototypical emotion such as anger, happiness and sadness in addition to the neutral state and in some cases even improve the classification that can be done directly by training basic emotions models from acoustic features. However, it is very important to emphasize that by using Emotion Primitives we can generalize the detection of emotional states in a much broader sense because we can define different areas in the three-dimensional space depending on the emotional states of interest according to the application and the emotional states that

Table 10
Accuracy in classifying Emotion Primitives by high and low classes.

Emotion Primitive	Accuracy
Valence	76.808
Activation	83.5411
Dominance	88.0299

Table 11
Recall when classifying basic emotions from Emotion Primitives.

Recall	Anger	Happiness	Neutral	Sadness
Continuous	85.82	54.93	79.17	42.11
Discrete	74.92	43.52	75.96	67.10
Baseline	69.68	21.01	35.23	76.84

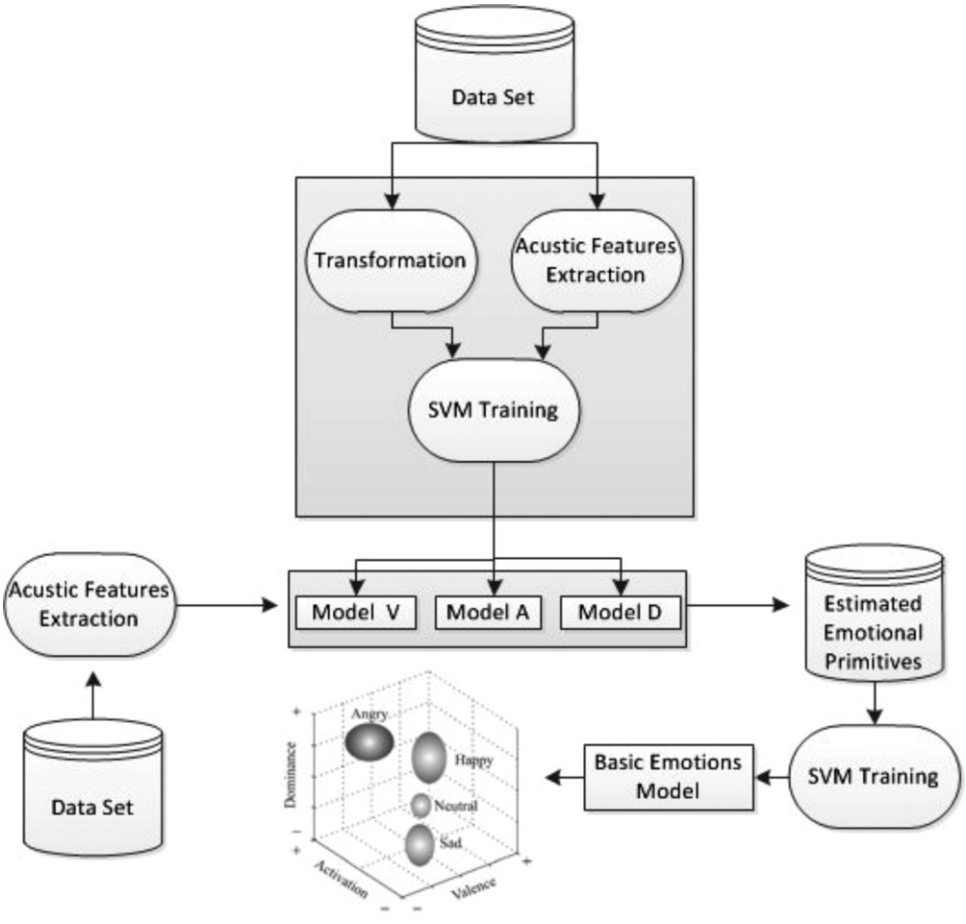

Fig. 4. Basic emotion classification process from emotion primitives.

are required to be discriminated. For example, an application might be interested in emotional states close to anger, or emotional states with high Activation, or negative Valence or any combination of the three primitives.

9. Conclusions

We carried out a study about the importance of different acoustic feature types from a continuous three-dimensional emotions model point of view. We analyzed each Emotion Primitive separately. Through the identification of the best features the automatic estimation of Emotion Primitives becomes more accurately and thus the recognition and classification of people's emotional state improves. To our knowledge the importance of acoustic features has not been studied from this approach. We have taken some ideas used in the study of the impact of features in the classification of discrete emotional categories, like share and portion metrics, and we have applied them to the continuous approach. We divided our 6920 features set into 12 groups according to their acoustic properties. We calculated some metrics for each group in order to estimate their performance in the automatic estimation of Emotion Primitives (correlation coefficient) and their contribution to the final set of features (share and portion). We worked with a database of acted emotions; labeled with the two most important annotation schemes, continuous and discrete. Despite being acted this database was designed trying to make it least artificial through the improvisation of dialogues.

We proposed two feature selection schemes taking into account that we had many features, but a very limited set of instances. We observed that both feature selection schemes obtained very similar results and they generally agree on the importance of feature groups.

The main contribution of this paper is the analysis of acoustic features. This analysis was based mainly on the correlation obtained by the models generated from a machine learning process (SVM). We realized that spectral feature groups are very important for the three primitives. According to our results the most important feature groups for each Emotion Primitive are:

- Valence: MEL – MFCC – cochleagrams – LPC
- Activation: MFCC – cochleagrams – energy – LPC
- Dominance: MFCC – cochleagrams – LPC – MEL

Clearly MFCC, LPC and cochleagrams groups are very important to estimate the three Emotion Primitives, since they appear among the most important for the three primitives. These three groups belong to the spectral information category, we can conclude from this fact that spectral analysis is more important than prosodic and voice quality analysis for Emotion Primitive estimation, except for Activation, where energy is also very important. Cochleagrams group is an interesting finding because, to our knowledge, it has not been used before for emotion recognition.

We observed that there were similar correlation results for selective and brute force feature extraction approaches when the selected features belonging to these sets were tested separately although the brute force set was much larger that the selective set. The portion for the selective was always much higher than portion for brute force, this tell us that the selective set has fewer features, but more important ones. The share was more balanced for both approaches, tending to be higher for brute force.

We proposed a way for mapping Emotion Primitives into basic emotions and we performed a basic emotion classification to show the feasibility of applying the continuous approach. The instance/attribute selection process based on Emotion Primitives has been successful in the discrimination of emotions reflected in the classification of basic emotions, raising the recall property for three out of four emotions studied in this work.

Interesting results have been obtained in this paper. However, a limitation is that experiments were done on an acted speech database. It is known that there are major differences between working with acted and real data. Another limitation is the fact that the recordings we used belong only to two different people. The next task will be to validate our findings with the extended version of the IEMOCAP database and with other databases, including different languages and different recording conditions. We also plan to refine the mapping and classification scheme, trying to take better advantage of continuous emotional model.

Acknowledgements

The authors wish to express their gratitude for the support given to carry out this research to the National Council of Science and Technology of Mexico through the postgraduate scholarship 49296 and the project 106013.

References

[1] C. Vera-Muñoz, L. Pastor-Sanz, G. Fico, M. Arredondo, A Wearable EMG Monitoring System for Emotions Assessment, first ed., Springer Publishing Company, Incorporated, 2008.
[2] G. González, Bilingual computer-assisted psychological assessment: an innovative approach for screening depression in chicanos/latinos, Tech. rep., University of Michigan, 1999.
[3] F. Nasoz, K. Alvarez, L. Lisetti, N. Finkelstein, Emotion recognition from physiological signals using wireless sensors for presence technologies, Cognition, Technology & Work 6 (2004) 4–14, URL: http://portal.acm.org/citation.cfm?id=1008243.1008250.
[4] L. Vidrascu, L. Devillers, Real-life emotion representation and detection in call centers data, in: ACII, 2005, pp. 739–746.
[5] M.S. Hussain, R.A. Calvo, A framework for multimodal affect recognition, in: HCSNet Perception and Action Workshop: Tools and Techniques for Conducting EEG and MEG Experiments, 2009, pp. 1–8.
[6] M. Murugappan, M. Rizon, R. Nagarajan, S. Yaacob, D. Hazry, I. Zunaidi, 4th Kuala Lumpur International Conference on Biomedical Engineering, Berlin, Heidelberg, 2008.
[7] H. Schlosberg, Three dimensions of emotion, Psychological Review 61 (2) (1954) 81–88.
[8] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, S.S. Narayanan, IEMOCAP: Interactive Emotional Dyadic Motion Capture Database, Journal of Language Resources and Evaluation 42 (4) (2008) 335–359.
[9] M. Lugger, B. Yang, Cascaded emotion classification via psychological emotion dimensions using large set of voice quality parameters, in: International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2008), Institute of Electrical and Electronics Engineers, 2008, pp. 4945–4948.
[10] M. Wollmer, F. Eyben, B. Schuller, E. Douglas-Cowie, R. Cowie, Data-driven clustering in emotional space for affect recognition using discriminatively trained lstm networks, in: Interspeech 2009, International Speech Communication Association, 2009, pp. 1595–1598.
[11] F. Eyben, M. Wollmer, A. Graves, B. Schuller, E. Douglas-Cowie, R. Cowie, On-line emotion recognition in a 3-d activation-valence-time continuum using acoustic and linguistic cues, Journal on Multimodal User Interfaces 3 (1–2) (2010) 7–19.
[12] B. Xie, L. Chen, G. Chen, C. Chen, Statistical Feature Selection for Mandarin Speech Emotion Recognition. Lecture Notes in Computer Science, Springer, Berlin/Heidelberg, 2005.
[13] M. Lugger, B. Yang, An incremental analysis of different feature groups in speaker independent emotion recognition, in: Proceedings of the International Conference on Phonetic Sciences, 2007, pp. 2149–2152.
[14] A. Batliner, S. Steidl, B. Schuller, D. Seppi, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, V. Aharonson, L. Kessous, N. Amir, Whodunnit – searching for the most important feature types signalling emotion-related user states in speech, Computer Speech and Language 25 (1) (2010) 4–28.
[15] S. Steidl, Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech, first ed., Logos Verlag, 2009, URL: http://www5.informatik.uni-erlangen.de/Forschung/Publikationen/2009/Steidl09-ACO.pdf.
[16] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, B. Weiss, A database of german emotional speech, in: Interspeech 2005, International Speech Communication Association, 2005, pp. 1517–1520.
[17] J. Montero, Estrategias para la mejora de la naturalidad y la incorporación de variedad emocional a la conversión texto a voz en castellano, Ph.D. thesis, Universidad Politécnica de Madrid, 2003.
[18] M. Grimm, K. Kroschel, S. Narayanan, The vera am mittag german audio–visual emotional speech database, in: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME 2008), 2008, pp. 865–868.

[19] P. Boersma, Praat, a system for doing phonetics by computer, Glot International 5 (2001) 341–345.
[20] F. Eyben, M. Wollmer, B. Schuller, Openear – introducing the munich open-source emotion and affect recognition toolkit, in: Proceedings of 4th International HUMAINE Association Conference on Affective Computing and Intelligent Interaction, 2009, pp. 1–6.
[21] H. Pérez-Espinosa, C.A. Reyes-García, Detection of negative emotional state in speech with anfis and genetic algorithms, in: Worshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA 2009), Firenze University Press, 2009, pp. 25–28.
[22] H. Pérez-Espinosa, C.A. Reyes-García, L. Villaseñor Pineda, Features selection for primitives estimation on emotional speech, in: International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2010), Institute of Electrical and Electronics Engineers, Dallas, Texas, 2010, pp. 5138–5141.
[23] K. Santiago, C.A. Reyes-García, M. Gomez, Conjuntos difusos tipo 2 aplicados a la comparación difusa de patrones para clasificación de llanto de infantes con riesgo neurológico, Master's thesis, INAOE, Tonantzintla, Puebla, México, 2009.
[24] A.L. Reyes, Un método para la identificación del lenguaje hablado utilizando información suprasegmental, Ph.D. thesis, INAOE, Tonantzintla, Puebla, México, 2007.
[25] T. Dubuisson, T. Dutoit, B. Gosselin, M. Remacle, On the use of the correlation between acoustic descriptors for the normal/pathological voices discrimination, EURASIP Journal on Advances in Signal Processing, Analysis and Signal Processing of Oesophageal and Pathological Voices 10.1155/2009/173967.
[26] C.T. Ishi, H. Ishiguro, N. Hagita, Proposal of acoustic measures for automatic detection of vocal fry, in: Interspeech 2005, International Speech Communication Association, 2005, pp. 481–484.
[27] R. Kehrein, The prosody of authentic emotions, in: Speech Prosody 2002, International Conference, 2002, pp. 423–426.
[28] M. Lugger, B. Yang, Classification of different speaking groups by means of voice quality parameters, ITG-Fachtagung Sprach-Kommunikation.
[29] B.F. Nuñes, Evaluación perceptual de la disfonía: correlación con los parametros acústicos y fiabilidad, Acta otorrinolaringológica española: Organo oficial de la Sociedad española de otorrinolaringología y patología cérvico-facial 55 (6) (2004) 282–287.
[30] A.B. Kandali, A. Routray, T.K. Basu, Vocal emotion recognition in five native languages of assam using new wavelet features, International Journal of Speech Technology 12 (1) (2009) 1–13.
[31] W. Byrne, J. Robinson, S. Shamma, The auditory processing and recognition of speech, in: Proceedings of the Workshop on Speech and Natural Language, HLT'89, Association for Computational Linguistics, Stroudsburg, PA, USA, 1989, pp. 325–331, URL: http://dx.doi.org/10.3115/1075434.1075490.
[32] T. Zbynik, J. Psutka, Speech production based on the mel-frequency cepstral coefficients, in: Eurospeech 1999, International Speech Communication Association, 1999, pp. 2335–2338.
[33] P. Dumouchel, N. Dehak, R. Attabi, Y.B.N Dehak, Cepstral and long-term features for emotion recognition, in: Interspeech 2009, International Speech Communication Association, 2009.
[34] E. Bozkurt, Improving automatic emotion recognition from speech signals, in: Interspeech 2009, International Speech Communication Association, 2009.
[35] B. Vlasenko, B. Schuller, A. Wendemuth, G. Rigoll, Frame vs. turn-level: emotion recognition from speech considering static and dynamic processing, in: A. Paiva, R. Prada, R. Picard (Eds.), Affective Computing and Intelligent Interaction, Lecture Notes in Computer Science, vol. 4738, Springer, Berlin/Heidelberg, 2007, pp. 139–147.
[36] O. Fujimura, K. Honda, H. Kawahara, Y. Konparu, M. Morise, J.C. Williams, Noh voice quality, Logoped Phoniatr Vocol 34 (4) (2009) 157–170.
[37] A. Metallinou, S. Lee, S.S. Narayanan, Decision level combination of multiple modalities for recognition and analysis of emotional expression, in: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2010), Institute of Electrical and Electronics Engineers, 2010, pp. 2462–2465.
[38] I.H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. (The Morgan Kaufmann Series in Data Management Systems), first ed., Morgan Kaufmann, 1999, URL: http://www.worldcat.org/isbn/1558605525.
[39] P. Pudil, J. Novovičová, J. Kittler, Floating search methods in feature selection, Pattern Recognition Letters 15 (1994) 1119–1125.
[40] M. Gutlein, E. Frank, M. Hall, A. Karwath, Large-scale attribute selection using wrappers, in: IEEE Symposium on Computational Intelligence and Data Mining 2009, Institute of Electrical and Electronics Engineers, 2009, pp. 332–339.