A Hybrid Distance-based Method and Support Vector Machines for Emotional Speech Detection

Vladimer Kobayashi

Université Jean Monnet
Laboratoire Hubert Curien CNRS, UMR 5516
18 rue du Pr. Benoit Lauras, 42000 Saint-Etienne, France
vladimer.kobayashi@univ-st-etienne.fr

Abstract. We describe a novel methodology that is applicable in the detection of emotions from speech signals. The methodology is useful if we can safely ignore sequence information since it constructs static feature vectors to represent a sequence of values; this is the case of the current application. In the initial feature extraction part, the speech signals are cut into speech segments of specified duration. The speech segments are processed and described using features such as pitch, energy, mel frequency cepstrum coeffiecients and linear prediction cepstrum coefficients. Our proposed methodology consists of two steps. The first step constructs emotion models using principal component analysis and it computes distances of the observations to each emotion models. The distance values from the previous step are used to train a support vector machine classifier that can identify the affective content of a speech signal. We note that our method is not only applicable for speech signal, it can also be used to analyse other data of similar nature. The proposed method is tested using two emotional databases. Results showed competitive performance yielding an average accuracy of greater than 90% on both databases for the detection of three emotions.

Key words: emotion recognition from speech, support vector machines, speech segment-level analysis

1 Introduction

The advent of Ubiquitous Computing research has paved the way to the development of systems that are more accustomed and able to respond in a timely manner according to human needs and behaviour [1, 2]. Computers and other machines have been successfully integrated to every aspect of life. To be truly practical machines must not only "think" but also "feel" since meaningful experiences are communicated through changes in affective states or emotions. At the heart of all of this is the type of data that we will handle to proceed with the computational task. In a typical scenario we deal with data types commonly encountered in signal processing since emotions are either overtly expressed in voice signals or covertly in biological signals and these signals can be captured

through the use of sophisticated sensors. The challenge not only lies on the pre-processing part but also on the development of methods adapted to the nature of data we wish to analyse.

In the past decade we have seen the explosion of studies that try to deduce human emotions from various signals we collect. By far the most carefully studied signal for this purpose is the speech signal [3, 4]. It is an accepted fact that it is relatively easy for us humans to identify (to a certain degree) the emotion of another person based on his voice, although, we are still trying to understand how we manage to do it. There are features in speech signals which we unknowingly distinguish and process that enable us to detect certain types of emotions. The central idea of many researches is to automate the process of detecting emotions which is vital to the creation of emotion-aware technologies and systems [4].

In this paper, as a first step, we also deal with speech signals. We argue that speech signals are the most convenient and reasonable to deal with since in real setting they can be easily captured. Unlike other signals such as ECG and EEG, speech signals are not particularly troublesome to collect and can be recorded anywhere and at any time. Also, many studies were published about processing of speech signals thus we can try the features proposed in previous studies in this work. The true nature of a speech signal is dynamic, a sequence of values indexed by time, however, the approach that we follow is to represent it as a single static feature vector. This approach is influenced by the fact that the sequence information is not essential for emotion recognition.

The objective of this work is the proposal of a novel technique that predicts emotions from speech signal. In essence our proposed method consisted of two steps. The first step is the creation of *emotion models* and the computation of deviations or distances of the speech signal "parts" from each of the emotion models. The second stage is to use the distance values computed in the previous step to construct a classifier that can detect the over-all emotional content of a given speech signal. In contrast with other techniques, we obtained additional knowledge such as the importance of different variables and the degree of separation of the speech signal components from each emotion categories. The first step is reminiscent of the method called Soft Modelling of Class Analogies (SIMCA) [5] although we do not attempt to classify the speech signals in this step. Another advantage of our technique is the tremendous flexibility it offers. Among other things, the modeller has the freedom to use different sets of features in both the first and second steps and adjust the underlying methods to make it robust to noise.

As a primary application of our proposed approach we deal with the problem of detecting three types of emotions, namely, "Disgust", "Angry", and "Sadness". These emotions are commonly encountered in application such as the analysis of telephone conversation and diagnosis of certain medical disorders. We tested our approach on two emotional speech databases. The results showed the effectiveness of our approach based from the analysis using the two databases. The accuracies are at least 90% on both databases.

The rest of the paper is organized as follows. Section 2 provides description of the two speech databases to which we tested our approach. Section 3 discusses the pre-processing and the extraction of speech acoustic features. Section 4 elaborates on our proposed approach. The results of our experiments are presented in Section 5. Finally, we close the paper in Section 6.

2 Speech Databases

We tested our proposed approach using the Berlin Database of Emotional Speech and the RML Database.

The **Berlin Database of Emotional Speech**[1] [6], also known as **Berlin EmoDB**, has been utilized in several studies as a benchmark database to test the performance of a technique for speech emotion recognition. The database was constructed using 10 speakers, 5 males and 5 females, who read 10 sentences in German that have little emotional content textually but they are read in such a way to simulate emotions. Originally there were seven discrete emotions considered. The sentence utterances are of variable lengths ranging from 1 to 4 seconds. There is a total of 535 sentence utterances that were evaluated and labelled in a perception test. We only retained those utterances which have emotion labels "Anger", "Disgust", and "Sadness". The recording was done in a studio with high-quality recording apparatus and stored in a mono wave file format with sample rate of 16,000 Hz and quantitative bits of 16.

The **RML Emotion Database**[2] contains 720 audio-visual emotional expression samples that were collected at Ryerson Multimedia Lab. Among the emotions that were expressed we only used "Anger", "Disgust", and "Sadness". This database is language and cultural background independent. The video samples were collected from eight human subjects, speaking six different languages (English, Mandarin, Urdu, Punjabi, Persian, Italian). Different accents of English and Chinese were also included. The samples were recorded at a sampling rate of 22050 Hz using a single channel 16-bit digitization. Samples have length ranging from 3 to 6 seconds with emotions being expressed.

For the two databases we consider sentence utterances as our speech signals.

3 Pre-processing and Feature Extraction

Instead of dealing with sequence of values and taking into account the dynamic nature of speech signal we took a slightly different approach. We firstly cut each speech signal into segments of 250 milliseconds (ms) each and we represent each segment by a single static feature vector. We found out that a length of 250 ms achieves good trade-off between emotional content and variability. In this part, features do not directly describe the whole speech signal but rather they describe the individual segments, thus a speech segment becomes our unit of analysis.

[1] `http://database.syntheticspeech.de/`
[2] `http://www.rml.ryerson.ca/rml-emotion-database.html`

3.1 Pre-processing

The speech signals were initially preprocessed by removing the silent parts at the beginning and end of the signals. Then they were cut into non-overlapping segments of 250 ms each. Here, we did a blind segmentation approach where no prior delimitation of word or syllable boundary has been performed. We assumed that segments may contain emotional primitives that contribute to the over-all emotional content of a speech signal. The segmentation as well as the subsequent extraction of features are illustrated in Figure 1

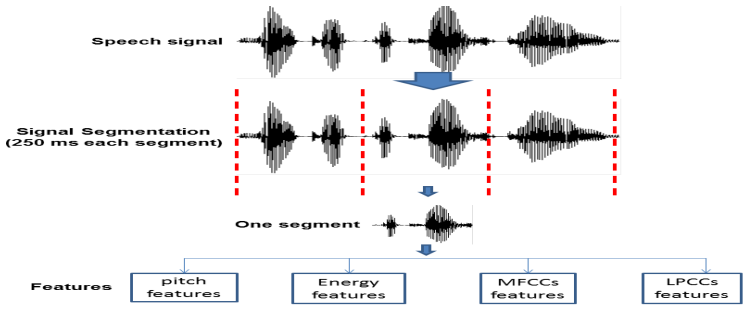

Fig. 1. Segmentation and Feature Extraction Part. An utterance is segmented into non-overlapping speech segments of 250 ms each. The features at this point are extracted from the individual speech segments.

3.2 Speech Acoustic Features

Here the unit of analysis is not the utterance but the 250 ms speech segments cut from it. To extract the segment-level features we apply short time analysis. From the name itself, short time analysis extract short term features on a frame basis; the reason why this type of analysis is also called frame-based analysis. We start by first decomposing the segments into a series of overlapping frames of specified duration. Here the duration of each frame is 25 ms and obtained every 10 ms using a Hamming window function. Four groups of feature types were considered: pitch, energy, mel-frequency cepstrum coefficients, and linear prediction cepstrum coefficients. These acoustic features have been demonstrated to be useful indicators of some emotions in speech signals.

The tension of the vocal folds and the sub glottal air pressure partially reflect the type of emotion expressed in speech. Pitch signal is produced from the vibration of the vocal folds and its vibration rate is called the fundamental frequency of the phonation F0 or pitch frequency. A lot of algorithms exist to estimate the pitch frequency. In this study we used the algorithm based on the autocorrelation of center-clipped frames. For each frame we computed the

value of pitch frequency and statistics of pitch are obtained for the entire segment. The computed statistics are the minimum, maximum, median, lower- and upper- hinges.

Energy is a basic feature in speech signal and it is related to the arousal level of emotions. We extracted the short-term energy on each frame. For the whole segment we computed the minimum, maximum, median, lower- and upper-hinges of the energies as features.

The **mel-frequency cepstrum coefficients** (MFCCs) are based on the characteristics of the human ear's hearing, which uses a non-linear frequency unit to simulate the human auditory system. It is the most widely used spectral representation of speech in many applications. Among their many advantages, MFCCs are simple to calculate, good ability of distinction, and anti-noise. For each of the frames, twelve standard MFCCs are calculated by following the steps: (1) taking the absolute value of the short time Fourier transform (STFT), (2) warping it to a Mel frequency scale, (3) taking the logarithms at each of the mel frequencies (4) taking the discrete cosine transform (DCT) of the log-Mel spectrum and (5) returning the first twelve components

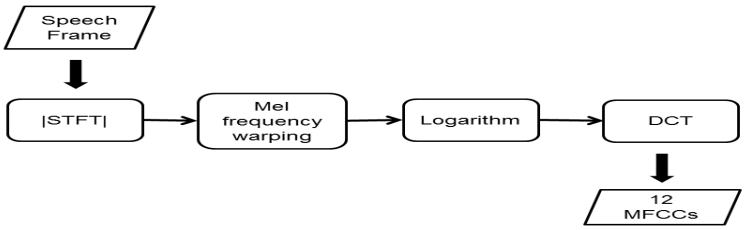

Fig. 2. Diagrammatic representation of the extraction of MFCCs

Figure 2 illustrate the whole process of obtaining the MFCCs. To get the MFCCs features for the whole segment we computed the mean of each of the twelve coefficients from among the constituent frames.

Lastly, **linear prediction cepstrum coefficients** (LPCCs) embody the characteristics of particular channel of speech. The same person expressing different emotions in speech will have different channel characteristics, thus, LPCCs are good to identify emotional content of speech. They are computed from linear prediction coding (LPC) coefficients using a recursive algorithm. As our features we extracted the cepstrum coefficients from 40 LPC coefficients from which we obtained 41 coefficients. The mean of each of the 41 coefficients from frames are calculated to represent the entire segment.

We refer the reader to [7] for additional information regarding the extraction of the preceding acoustic features. A total of 63 features (5 pitch related statistics, 5 energy related statistics, 12 MFCCs and 41 LPCCs) were considered in this study. The features discussed here are used in the first step of our proposed approach.

4 Our Proposal

Our proposed approach does not only achieve maximum detection rate for the three types of emotion but also it provides added information toward the understanding of the synergy among emotions within a speech signal. By doing this, we are able to make a detailed analysis of a speech signal by examining its speech segment components. The characterisation of speech segments could give us a complete emotional expression of the whole signal. In other words, by examining the parts we understand the whole.

Our proposal consisted of two steps. The whole process is shown in Figure 3. The first step involved building models for each emotion class. Also included in this step is the computation of distance measures which will quantify the deviations of the speech frames to each emotion class. The second step will involved the construction of the second set of features by computing summary statistics of the distance values and the training of speech signal classifier.

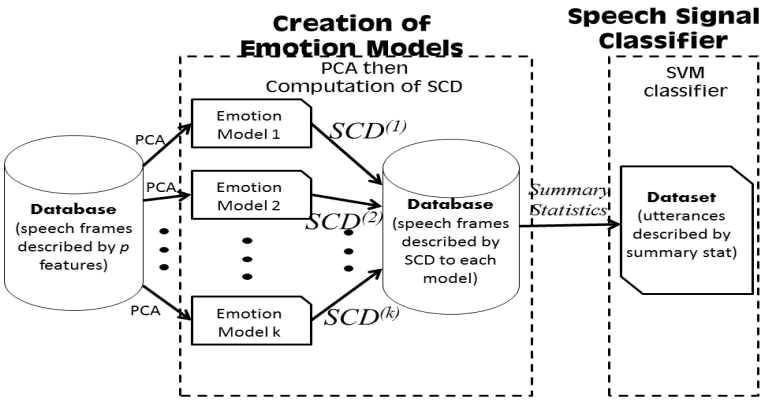

Fig. 3. Diagrammatic presentation of our proposed approach

4.1 Preliminaries

Let us denote an emotion category (or emotion model) as E^j, for 3 emotions we have $j = 1, 2, 3$. Remember that a speech signal (mother signal) is cut into speech segments (or simply segments) of 250 ms duration. We do this because in actual conversation speech signals come in continuous form and not per utterance with clear boundaries. Furthermore, this course of action saves us from the unnecessary computational step of identifying word or syllable boundaries to compute the features. This makes our approach more practical since the scheme is suitable for real-time processing and adaptable to stream analysis.

Depending on the length of the speech signal the number of segments may vary. We assigned emotion labels for each of the segments during training. Here we assume that *the emotion label of a segment is the same as the emotion label of the mother signal where that segment came from.* We represent a segment as $\mathbf{s}_i^j$, which can be interpreted as segment i that has emotion label E^j. Finally each segment is described by p (=63) features. Thus for a given segment its static single feature vector representation is $\mathbf{s}_i^j = (s_{i1}^j, s_{i2}^j, \ldots, s_{ip}^j)^T$. This part is a good reference if ever you need to refresh your memory regarding certain notations.

4.2 Construction of Emotion Models

The motivation for this step is to reveal underlying structure for each emotion categories. This way we understand better the characteristics of speech frames in each emotion categories. Also, this stage will involve a feature reduction part since we want to derive features that are really useful to describe each group.

To obtain the emotion models, we run Principal Component Analysis (PCA) on each emotion classes. After we run PCA on each group E^j we will obtain a matrix of scores T^j and loadings P^j for each group. Since we run separate PCA on each emotion classes we can now summarize each emotion classes in different subspace models (according to the PCA models). The number of retained principal components for each class is denoted by $k_j \ll p$. The relevance of the features on each model can be assessed by examining the loadings of the features on the extracted principal components.

Once we have the emotion models, we can now compute the deviations of each segments to the different models. We can define two deviations: the *orthogonal distance* (OD) and the *score distance* (SD).

The orthogonal distance is simply the Euclidean distance of a segment to an emotion model in its PCA subspace. To compute the orthogonal distance of any segment $\mathbf{s}$, we first compute its projection $\mathbf{s}^{(j)}$ on emotion model E^j. Its projection is given by

$$\mathbf{s}^{(j)} = \bar{\mathbf{s}}^j + P^j (P^j)^T (\mathbf{s} - \bar{\mathbf{s}}^j) \tag{1}$$

where $\bar{\mathbf{s}}^j$ is the feature vector mean (column means) of the speech frames in group E^j. The OD to group E^j of the segment $\mathbf{s}$ is defined as the norm of its deviation from its projection, specifically,

$$\mathrm{OD}^{(j)} = \|\mathbf{s} - \mathbf{s}^{(j)}\| \tag{2}$$

On the other hand, the score distance is a robust version of the Mahalanobis distance measured in PCA subspace. Hence, the score distance of $\mathbf{s}$ is provided by:

$$\mathrm{SD}^{(j)} = \sqrt{(\mathbf{t}^{(j)})^T J^{-1} \mathbf{t}^{(j)}} = \sqrt{\sum_{a=1}^{k_j} \frac{(t_a^{(j)})^2}{\lambda_a^{(j)}}} \tag{3}$$

where $\mathbf{t}^{(j)} = (P^j)^T(\mathbf{s} - \bar{\mathbf{s}}^j) = (t_1^{(j)}, t_2^{(j)}, \ldots, t_{k_j}^{(j)})^T$ is the score of $\mathbf{s}$ with respect to the E^j group, $\lambda_a^{(j)}$ for $a = 1, 2, \ldots, k_j$ stands for the largest eigenvalues in the E^j group, and J is the diagonal matrix of the eigenvalues. The advantage of SD over OD is that SD uses information about the eigenvalues.

In the usual case we need to decide which of the two distances is more appropriate for the problem, but we opted to combine the two distances by using a parameter γ. Specifically we define the combined distance and named it *scortho distance* (SCD) for given $\mathbf{s}$ by

$$\text{SCD}^{(j)} = (\gamma)\text{OD}^{(j)} + (1 - \gamma)\text{SD}^{(j)} \tag{4}$$

where $\gamma \in [0, 1]$. We can optimized the results by choosing appropriate values for the new parameter. One way to choose the parameter is to perform cross-validation and optimizing certain criterion like test prediction accuracy.

Another advantage of building emotion models is we can identify segments which are markedly far from any of the models. We do this by identifying a cut-off value in the computed distances. The cut-off values can be computed by examining the distribution of the distance values. For example, in the case for the score distance, the squared score distances follow asymptotically a χ^2-distribution with k_j degrees of freedom thus we can set $c_{SD}^j = \sqrt{\chi^2_{k_j;0.975}}$ as the cut-off value. We can perform the same kind of analysis for the orthogonal distance and the scortho distance. Segment which computed distances are outside the cut-off values for all the models with respect to a particular distance are termed unrepresentative segments because purportedly they do not contain any emotion. Another possibility is they may form an unknown group and after detection can lead to a new knowledge about the nature of the problem we are studying. In this paper, we are not yet going to pursue the idea of identifying unrepresentative segments.

To summarize the first step of our proposed approach, we first build emotion models using PCA and then proceed to the computation of the SCD. At this point, each speech frame is represented by a vector consisting of its distances to each emotion models. Thus, in our case since we have three emotions, the speech frames are now represented by a vector of three components. Notationally, using the SCD distances, given $\mathbf{s}$, we have

$$\mathbf{s} = (\text{SCD}^{(1)}, \text{SCD}^{(2)}, \text{SCD}^{(3)}) \tag{5}$$

This new representation of the speech frames will be used in the second step.

4.3 Speech Signal Level Classifier

The new representation of the segments obtained from the previous step is aggregated in the speech signal level. Remember that we cut the speech signals into segments so that we can proceed with the initial feature extraction. The aggregation is made possible by computing some summary statistics. Suppose we have a speech signal u and the speech frames cut from it are $\{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_m\}$.

One summary statistic that we can compute is the mean. Hence, using the mean we can represent the speech signal u as

$$\mathbf{u} = (\mu_1, \mu_2, \ldots, \mu_l) \tag{6}$$

where l is the number of emotion categories and

$$\mu_j = \frac{1}{m} \sum_{i=1}^{m} \mathrm{SCD}_i^{(j)} \tag{7}$$

where, $\mathrm{SCD}_i^{(j)}$ is the distance of ith segment extracted from u to emotion model E^j. The interpretation of the components is straightforward in this case: the μ_j are the mean SCD of the speech frames components in u to emotion model E^j. We can also view this as the "distance" of our speech signal to each emotion categories.

Aside from the mean, we can also use additional summary statistics like the standard deviation or the inter-quartile range to capture additional information. It is important to note that if we use many statistical measures we will be increasing the size of the vector representation of the speech signals. For instance in our case with three emotions, if we consider two summary statistics then the number of vector components will be equal to 6. Thus it is imperative to choose the right summary statistics to better capture the important characteristics of the speech signals as expressed by the deviations of its member speech segments.

Once we have represented each speech signal u as feature vector $\mathbf{u}$ in the manner we described above we can now construct a speech signal matrix $\mathbf{U}$ denoted by

$$\mathbf{U} = \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_n \end{pmatrix} \tag{8}$$

where n is the total number of utterances or speech signals.

With emotion labels associated to each utterance we are now ready to train the classifier at this step. The user has the liberty to select the classification algorithm he will use here. For our case, we decided to employ Support Vector Machines since it has shown competitive performance in other pattern recognition problems [8, 9]. For an in-depth discussion about the principles of SVM we refer the reader to [10].

5 Results

All throughout the experiments we made use of the features in the speech frame level, namely, 5 pitch related statistics, 5 energy related statistics, 12 *MFCCs*, and 41 *LPCCs* in the first step. In the second step we have aggregated the distances obtained from the first step by computing the median, standard deviation,

and the inter-quartile range of the SCD distances of the speech frames within an utterance with respect to each emotion models. Thus, each speech signal is represented by a vector of 9 elements. Morever, we trained SVM classifier in the second step using the ordinary dot product kernel function. As a standard practice to get reliable estimate of the performance of our proposed approach we use 10-fold cross validation and computed the accuracies and macro F-measure using the test data for each fold. It is important to emphasize here that we trained the models without taking into account the gender or age or language of speakers and the results reported are the mean accuracies and mean macro F-measures on the speech signal level. The whole process of train and testing is depicted in Figure 4. We applied our approach on the two databases separately.

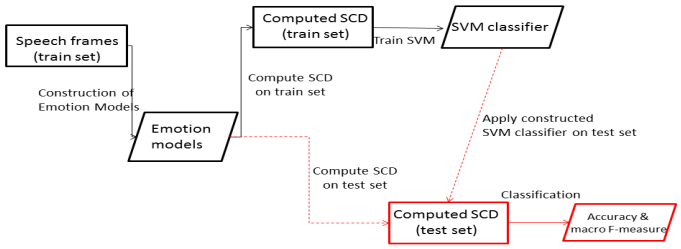

Fig. 4. Training and testing followed in the experiments

Table 1. Average precision and recall for each emotion categories in each database using our proposed approach

Database	Emotion	Ave. Precision	Ave. Recall
Berlin EmoDB	Disgust	0.907	0.800
	Sadness	0.967	1.000
	Anger	0.955	0.95
RML	Disgust	0.850	0.900
	Sadness	1.000	1.000
	Anger	0.975	0.95

From the results shown in Table 1 we find that our approach is able to effectively identify the three types of emotions. Particularly our approach is superior in detecting emotions “Sadness” and “Anger”. This can be explained by the fact that the two emotions have contrasting arousal characteristics as expressed in the speech signal. “Sadness” is a lowly active emotion and “Anger” is a highly-active one. The case of emotion “Disgust” is interesting because

although it has almost the same arousal as "Anger", the classifier has successfully distinguish it from "Anger" most of the time especially in the RML database.

We present in Table 2 the mean accuracy and mean macro F-measures obtained by our approach. The table confirms our claim that our method is particularly effective in the detection of these three emotions.

Table 2. Performance of our proposed approach on the two databases assessed using average accuracy and average macro F-measure. We also present the baseline accuracy obtained by classifying all speech signals to only one emotion.

Database	Baseline Accu.	Ave. Accu.	Ave. macro F-measure
Berlin EmoDB	33.33%	92.10%	94.10%
RML	33.33%	95.70%	94.60%

Aside from the good detection rates we also get additional knowledge regarding the characteristics of the individual emotion classes. The information is derived from the first step of our methodology. We present in Table 3 the summary of the attributes of the emotion models constructed using the Berlin EmoDB. We find that although we have used an initial 63 features we discovered that we can construct at most 10 (latent) features and still capture the over-all variation in each emotion models. An analysis on the loadings of the original features to the constructed features revealed that energy related features and MFCCs are influential in the detection of "Disgust" and "Sadness" whereas pitch related features, energy related features, and MFCCs are important features for the detection of "Angry". LPCCs seem to have less contribution in this case. Maybe, LPCCs will become necessary if we wish detect other types of emotions.

Table 3. Summary characteristics of each emotion models for the Berlin EmoDB.

	Disgust Model	Sadness Model	Angry Model
PCs retained (latent features)	10	8	10
Explained Variance	100%	100%	100%
Most Important Variables	energy, MFCCs	energy, MFCCs	pitch, energy, MFCCs

6 Summary and Conclusion

The methodology we have proposed in this paper has shown competitive performance in the detection of three emotion types on the two databases. Aside from the good classification accuracies the methods also reveal additional knowledge regarding the individual emotion classes. This knowledge is desirable if we want

to understand better the type of features useful for the discrimination of emotion classes and also the synergy among emotion categories within speech signals.

Another confirmation we got in this study is the usefulness of using speech segments as unit of our analysis. The speech segments acts as atoms of emotions from which we can identify emotional primitives that could be used to deduce the over-all emotion of a speech signal.

In practice, our proposed approach can be implemented relatively fast. In the first step PCA can be run rapidly and the computation of distances is speedy. In the second step the complexity only depends on the complexity of the classifier. Lastly, our proposed approach offer new insights to the kind of analysis that can be done and additional tool to analyse emotional speech database.

In our future work we will investigate further the use of other features both in the first and second steps and classifiers in the second step. We will also test our method on other speech databases and with more number of emotions.

References

1. Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G.N., Kollias, S.D., Fellenz, W.A., Taylor, J.G.: Emotion recognition in human-computer interaction. IEEE Signal Processing Magazine **18** (2001) 32–80
2. Vogt, T., André, E., Wagner, J.: Automatic recognition of emotions from speech: A review of the literature and recommendations for practical realisation. In Peter, C., Beale, R., eds.: Affect and Emotion in Human-Computer Interaction, From Theory to Applications. Volume 4868 of LNCS. Springer (2008) 75–91
3. El Ayadi, M.M.H., Kamel, M.S., Karray, F.: Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition **44** (2011) 572–587
4. Koolagudi, S.G., Rao, K.S.: Emotion recognition from speech: a review. International Journal of Speech Technology **15** (2012) 99–117
5. Branden, K.V., Hubert, M.: Robust classification in high dimensions based on the SIMCA method. Chemometrics and Intelligent Laboratory Systems **79** (2005) 10–21
6. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W.F., Weiss, B.: A database of german emotional speech. In: INTERSPEECH 2005, ISCA (2005) 1517–1520
7. Rabiner, L., Schafer, R.: Introduction to Digital Speech Processing. Foundations and Trends in Technology. Now Publishers (2007)
8. Osuna, E., Freund, R., Girosi, F.: Training support vector machines: an application to face detection. In: CVPR '97, IEEE Computer Society (1997) 130–136
9. Joachims, T.: Text categorization with suport vector machines: Learning with many relevant features. In Nedellec, C., Rouveirol, C., eds.: Proc. 10th European Conference on Machine Learning. Volume 1398 of LNCS., Springer (1998) 137–142
10. Herbrich, R.: Learning Kernel Classifiers: Theory and Algorithms. MIT Press, Cambridge, MA, USA (2001)