Modeling the Temporal Evolution of Acoustic Parameters for Speech Emotion Recognition

Stavros Ntalampiras and Nikos Fakotakis

Abstract—During recent years, the field of emotional content analysis of speech signals has been gaining a lot of attention and several frameworks have been constructed by different researchers for recognition of human emotions in spoken utterances. This paper describes a series of exhaustive experiments which demonstrate the feasibility of recognizing human emotional states via integrating low level descriptors. Our aim is to investigate three different methodologies for integrating subsequent feature values. More specifically, we used the following methods: 1) short-term statistics, 2) spectral moments, and 3) autoregressive models. Additionally, we employed a newly introduced group of parameters which is based on the wavelet decomposition. These are compared with a baseline set comprised of descriptors which are usually used for the specific task. Subsequently, we experimented on fusing these sets on the feature and log-likelihood levels. The classification step is based on hidden Markov models, while several algorithms which can handle redundant information were used during fusion. We report results on the well-known and freely available database BERLIN using data of six emotional states. Our experiments show the importance of including information which is captured by the set based on multiresolution analysis and the efficacy of merging subsequent feature values.

Index Terms—Acoustic signal processing, speech emotion recognition, temporal feature integration, autoregressive models, wavelet decomposition.

1 Introduction

Emotion comprises one of the most basic factors with respect to the communication between humans. It would be ideal to have human emotions automatically recognized by machines, mainly for improving human-machine interaction. This motive is behind the constantly increasing attention that this particular scientific field has been receiving lately. A fruitful combination of completely different disciplines (such as psychology, neuroscience, and engineering) is essential in order to automatically categorize emotional content in speech signals. To construct such a framework, one mainly needs a well-annotated database, which may involve a psychologist as well as engineering skills (programming and signal processing). The basic difficulty is to cover the gap between the information which is captured by a microphone and the corresponding emotion, and to model the specific association. The particular gap can be very hard to bridge since a high level semantic interpretation of the audio signal is required which many times relies on difficult-to-capture aspects of the person under study, such as his cultural background, his ethnicity, etc. However, the problem is usually narrowed and deals with the following six emotions: happy, anger, fear, neutral, disgust, and sadness. It should be noted that the particular constraint does not compromise the applicability of the particular scientific domain on real-world problems. On the contrary, it is usual to first constrain a complicated problem toward providing a stable, efficient, and generic solution. Furthermore, a reduced space of emotions is adequate for most human computer interaction applications [1].

This work deals with the engineering part of the emotion recognition chain. The automatic detection of emotions in spoken language has become extremely important due to its immediate usage in many applications which include automatic dialog systems. For example, an automated speech-driven home assistant may update its functionality every time the emotions of the user change so as to provide enhanced services. This type of framework essentially characterizes a particular audio sequence using a label out of a set of predefined classes. This work proposes a generic methodology for emotion recognition after experimenting with new for the specific scientific field techniques. Various frameworks have been proposed in the literature during the recent past. The main emphasis is placed upon the database, the extracted features, and the pattern recognition algorithm. Our intention is not to provide an extensive overview of the related work, for which the interested reader may refer to [2], which is a well-structured and up to date survey. Most of the previous work is more or less laboratory-based systems, operating offline onto acted data sets. A great deal of feature sets have been used by the community which can be distinguished so far into the following four groups: *continuous* (e.g., energy and pitch), *qualitative* (e.g., voice quality), *spectral* (e.g., MFCC), and features based on the *Teager energy operator* (e.g., TEO autocorrelation envelope area [3]). With respect to the recognition phase a great deal of methodologies has been followed: HMM, GMM, ANN, *k*-NN, and several others as well as their combination which attempts to

- *The authors are with the Wire Communications Laboratory, Electrical and Computer Engineering Department, University of Patras, Sofocleous-Adiparou 1, 26500 Rion, Patras, Greece.*
E-mail: {sntalampiras, fakotaki}@upatras.gr.

Manuscript received 7 Mar. 2011; revised 10 Sept. 2011; accepted 13 Sept. 2011; published online 27 Sept. 2011.
Recommended for acceptance by S. Narayanan.
For information on obtaining reprints of this article, please send e-mail to: taffc@computer.org, and reference IEEECS Log Number TAFFC-2011-03-0019.
Digital Object Identifier no. 10.1109/T-AFFC.2011.31.

maintain the advantages of each classification technique. After studying the related literature we identified that the feature set which is mostly employed is comprised of pitch, MFCCs, and HNR. Additionally, the HMM technique is widely spread among the researchers due to its effectiveness and its general usage for a gamut of speech processing applications.

The closest papers to this work are [4], [5], [6], [7], and [8], where the temporal structure of the low level descriptors or a relatively large portion of the audio signal is taken into account in order to extract information which could be crucial during both the modeling and classification processes. Li and Zhao [4] computed speech features that represent entire utterances using the average values of various features which belong either to the time or to the frequency domain. For classifying the unknown recordings they used vector quantization, ANNs and GMMs. Their data set is comprised of speech, which is captured from student volunteers, and includes six emotions (neutral, happy, angry, fearful, surprised, and sad). Jiang and Cai [5] constructed an emotion recognition system that enables the combination of both statistic and temporal features. GMM and HMM likelihoods are integrated for forming a holistic description of each recording, which is then fed to a Bayesian and an MLP classifier. They experimented on a Chinese speech corpus of six emotions (anger, fear, happy, sad, surprise, and neutral) while their feature set was comprised of the F0, feature contours of log energy, and syllable duration. A different direction is followed by Vlasenko et al. [6], [7], where a variety of descriptors (MFCC, prosodic, speech quality, and articulatory) is computed on frame as well as turn level. During the frame level analysis they used a GMM classifier, while an SVM classifier was used with respect to the turn level. They conducted experiments using the BERLIN Emotional Speech database, as well as the Speech Under Simulated and Actual Stress (SUSAS), while the results are particularly good. The corresponding average recognition rates are 89.9 and 83.8 percent. Last but not least, an interesting approach is followed in [8], where features are derived from a spectro-temporal representation of speech. Their data are acquired from the BERLIN database and they reach 88.6 percent recognition accuracy using an SVM classifier with radial basis function. To summarize, the techniques which "merge" subsequent feature values have not been deeply exploited by the emotion recognition community, in contrast to other audio signal processing applications such as classification of musical genres [9], musical instrument recognition on solo musical phrases [10], and generalized sound recognition [11].

This work systematically explores different kinds of temporal feature integration techniques for classifying speech signals according to the emotion which is expressed. Three kinds of techniques are considered: 1) short term statistics, 2) spectral moments, and 3) autoregressive models (AR). Furthermore, we make use of a general purpose feature set which is derived from the wavelet domain. The particular domain has only been partially explored with respect to the general field of "affect" recognition and specifically stressed speech [12], [13]. In order to obtain comparative results we employed the BERLIN database including the "big six" emotional states: *neutral*, *happy*, *angry*, *fearful*, *surprised*, and *sad*. Following the recognition rates of every phase of our work we conducted extensive experiments with respect to fusing different feature sets and temporal integration methods at various levels. Our ultimate goal is to gain as much information as possible from the training data so as to form a generic and representative picture of the emotions included in the present study. In other words, we wish to study and understand the effect of temporal integration of acoustic parameters which belong to different domains—*frequency* and *wavelet*—for classifying human emotions.

The present paper is organized as follows: Sections 2 and 3 present the overall system architecture and the temporal integration methods, respectively, which were thoroughly investigated during our experimentations which are described in Sections 4 and 5. Section 4 provides the result analysis with respect to the early integration methods, while Section 5 is focused on the experimental results of the different fusion schemes (late integration). Our conclusions are drawn in the final section.

2 System Architecture for Recognizing Human Emotional States

This section analyzes the design and basic operational aspects of the emotion recognition system. Fig. 1 depicts both the training and testing processes. Initially, the audio signals are normalized by subtracting the mean value of each recording. Then, they are cut into overlapping frames where the feature extraction algorithms elaborate. This process ensures robustness to possible misalignments as the edges of sequential windows of raw data are overlapped. Furthermore, the windowing function enables smoothing of any discontinuities. Two feature sets are extracted: 1) The first one represents the baseline set, which includes widely used features with respect to the specific problem (Mel filterbank, pitch, and HNR) and 2) features from the wavelet domain calculated on specific frequency bands. These features are used both juxtaposed (feature-level fusion) and in parallel mode toward creating two different probabilistic models for each emotional class. Subsequently, these two feature sets are temporally integrated based on three different approaches: 1) statistical, 2) spectral, and 3) modeling using autoregressive processes. Thus, we construct probabilistic models for each integration methodology and for each feature set. The dimensionality of the temporally integrated feature sets is reduced using principal component analysis (PCA) technique. This phase serves two purposes: 1) keeping only a small number of the feature coefficients while their variance is maintained and 2) the decrement of computational complexity which is present during the model construction stage and needs special attention. PCA can efficiently maintain the variance of the data while using a relatively small number of coefficients. The above described approach is equivalent to identifying a *set* of uncorrelated acoustic parameters for emotion recognition, instead of just combining the best individual parameters. Afterward the distribution of the resulted feature values with respect to each emotional category is approximated by hidden Markov models. HMMs have the ability to

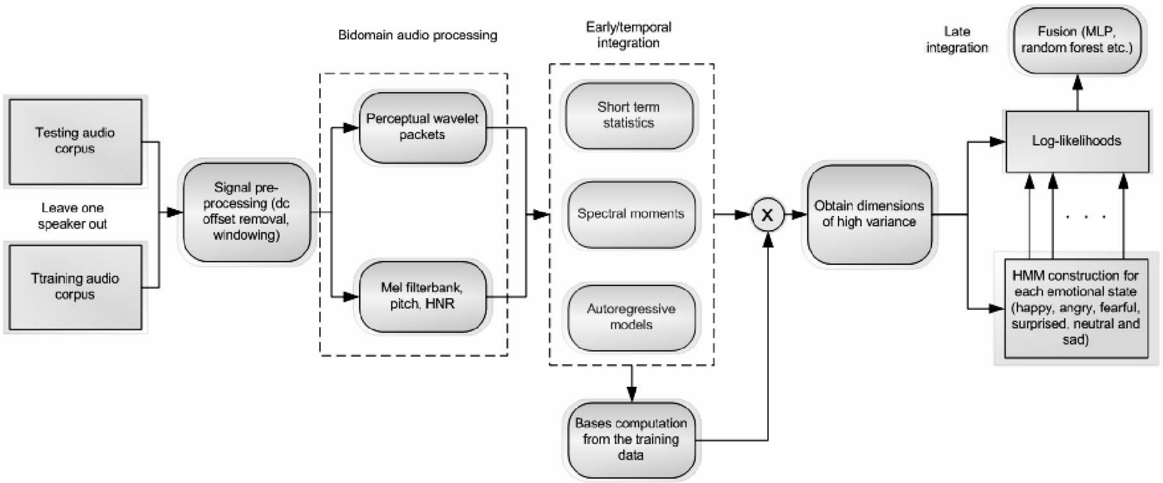
Bidomain audio processing
Early/temporal integration
Late integration
Fusion (MLP, random forest etc.)
Testing audio corpus
Leave one speaker out
Ttraining audio corpus
Signal pre-processing (dc offset removal, windowing)
Perceptual wavelet packets
Mel filterbank, pitch, HNR
Short term statistics
Spectral moments
Autoregressive models
X
Obtain dimensions of high variance
Log-likelihoods
. . .
HMM construction for each emotional state (happy, angry, fearful, surprised, neutral and sad)
Bases computation from the training data

Fig. 1. Overall architecture of the framework for automatic recognition of emotional states. The two integration phases are depicted: 1) early/feature-level and 2) late/log-likelihood level fusion.

model not only the static aspects of a feature sequence but also its temporal behavior. At the particular phase the class associated with the model which provided the highest log-likelihood is assigned to the novel audio recording. For improved recognition results, the final stage of the processing chain includes fusion of the log likelihoods which are produced by each modeling procedure. Fusion is conducted using various pattern recognition approaches, such as multilayer perceptron (MLP), random forest, etc. The following section explains the procedures which lead to the extraction of the two feature sets.

2.1 Analysis of the Feature Extraction Algorithms

Two feature sets of heterogeneous domains were employed in the current study: 1) the baseline set and 2) features based on multiresolution analysis. During their extraction the same parameters (frame and overlap sizes) were used so as to obtain comparable results for a thorough performance analysis. The first set includes Mel filterbank, pitch, and harmonic to noise ratio. For each frame of a given signal we compute 23 Mel filterbank log-energies. The logarithm is then calculated to space the data. These are appended to pitch and HNR, which is computed based on a forward cross-correlation analysis. Pitch and HNR are indicative of the variations that intonation exhibits among various emotional states. In addition, there is a strong association between certain pitch patterns and specific emotional states; thus the particular features should be useful for classification. The combination of Mel filterbank, pitch, and HNR is given as input to the temporal integration methodologies for forming the first group of descriptors.

The second set [11], [14] is designed to provide a complete analysis of an audio signal while different frequency bands are approximated by wavelet packets (WP). It takes into account that not all parts of the spectrum affect human perception in the same way. Consequently, the division of the spectrum requires a fine partitioning. In [15] and [16], it is observed that the human auditory system filters the entire audible spectrum into many critical bands. Based on this observation, we employed a critical-band-based filterbank with the frequency ranges denoted in Table 1 using Gabor bandpass filters. After the critical band-based analysis, we extract WPs of three levels. Downsampling is then applied following the Nyquist theorem. Subsequently, we compute the autocorrelation envelope area associated with each wavelet coefficient, which is then normalized by half the frame size. These comprise the perceptual wavelet packet (PWP) integration feature vector which is fed to the temporal integration methodologies. The corresponding block diagram is illustrated in Fig. 2. The specific feature set may capture the variations which are exhibited by each emotion within critical spectral areas while they are represented by wavelet coefficients. This information is crucial during both the modeling and the classification process.

2.2 PDF Estimation Using HMMs

The fundamental idea behind the recognition of every type of audio signals is that they follow a characteristic pattern for distributing their energy on their frequency content. We

TABLE 1
The Frequency Bands for
Perceptual Wavelet Packet Integration Analysis

Band Number	Lower (Hz)	Center (Hz)	Upper (Hz)
1	0	125	250
2	250	375	500
3	500	625	750
4	750	875	1000
5	1000	1125	1250
6	1250	1375	1500
7	1500	1625	1750
8	1750	1875	2000
9	2000	2250	2500
10	2500	2750	3000
11	3000	3250	3500
12	3500	3750	4000
13	4000	4250	4500
14	4500	4750	5000
15	5000	5500	6000
16	6000	6500	7000
17	7000	7500	8000

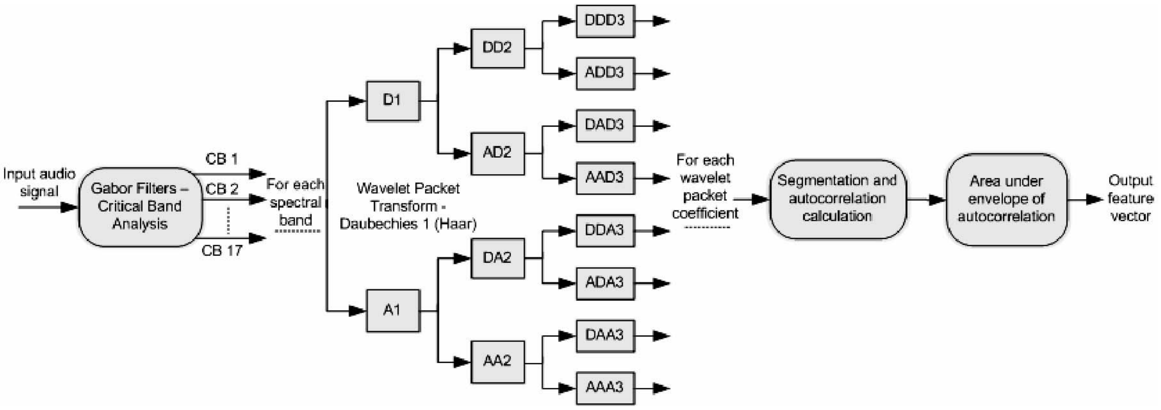
Input audio signal
Gabor Filters – Critical Band Analysis
CB 1
CB 2
CB 17
For each spectral band
Wavelet Packet Transform - Daubechies 1 (Haar)
D1
A1
DD2
AD2
DA2
AA2
DDD3
ADD3
DAD3
AAD3
DDA3
ADA3
DAA3
AAA3
For each wavelet packet coefficient
Segmentation and autocorrelation calculation
Area under envelope of autocorrelation
Output feature vector

Fig. 2. Critical-band-based audio analysis using wavelet packets for extracting the perceptual wavelet packet integration group of parameters.

try to capture this by selecting appropriate features and subsequently model their density based on the training data. The parameters of each model essentially represent the a priori knowledge that we have about the respective emotional category. In this paper, we adopt the HMM approach, which is frequently used by the emotion recognition community [2]. With this approach we break up the feature sequence into a predefined number of segments/states and the training algorithm learns the associations between them. We create a *kxk* transition matrix, wherein each element represents the probability of transition across different states, i.e., the element (i, j) is the probability of shifting to state j at time t given state i at time $t - 1$. In the left-right case which is followed here, there are no directed loops in the automation, while each state is modeled by a GMM with diagonal covariance matrix.

We used Torch [17] implementation of HMM, which is written in C++. The number of k-means iterations for initialization was 50, while the Baum-Welch algorithm had an upper limit of 25 iterations, with a threshold of 0.001 between subsequent iterations. We conducted extensive experiments with respect to

1. optimizing the statistical models in terms of number of states and Gaussian components,
2. exploring the performance of the feature sets,
3. investigating each temporal integration methodology, and
4. deciding upon the size of the integration length.

3 Temporal Feature Integration Methods

This section offers an insightful analysis of the temporal integration methods which were employed during the present study. Most of the previous work in the area of speech emotion recognition is focused on frame-based analysis, following the so-called bag-of-frames approach [18]. In many cases this kind of analysis has been proven effective [19], [20], but it would be of great interest to experiment toward creating a more compact representation of the feature sequence. This process may lead to a more characteristic picture of the signal under study since it would be focused on its global characteristics. This so-called *early* integration (as opposed to the late on which operates on the classifier level) is based on postprocessing low-level descriptors which are calculated on frames of relatively small duration, e.g., 30 ms. The idea behind temporal feature integration is to discard features which were computed on frames not representative of the emotion we try to model. It is usual that inside each sound recording there are parts which are not characteristic of the particular class. These segments are the ones which usually lead to misclassifications since they burden the system during both modeling and classification. By merging subsequent feature values we try to eliminate their influence on training/testing and putting the emphasis on the global characteristics of the specific class. As long as the integration methodology retains these characteristics the recognition capabilities of the system should be improved. The within class variability is reduced, which results in finer modeling of the shared characteristics among the samples of the same category of emotion. Our experiments are concentrated on finding the optimal value of the frames which are to be integrated (for both feature sets) as well as deciding on the integration strategy which provides the highest recognition accuracy.

Each integration function is applicable to a predefined number of frames and transforms them according to the following equation:

$$X_k = F(x_t, \ldots, x_{t+p-1}), \qquad (1)$$

where X_k denotes the integrated vector of the kth texture window (consecutive frames starting from frame t and ending at frame $t + p - 1$) and x_t is the value of feature x at frame t. The number of frames over which the integration is applied is called the *texture window* and is denoted as p. This equation provides a higher level description of the feature series. Several integration strategies are based on the computation of statistics over the texture window. Other strategies are based on the assumption that the feature sequence can be viewed as a random process (e.g., autoregressive models). The three different integration strategies which are investigated in this work are explained next.

3.1 Short-Term Statistics

The simplest way to perform temporal integration of a feature sequence is to compute its statistical moments. The aim here is to derive an accurate description of a specific texture window p with only a small number of its respective statistical measurements. The following five statistical moments are considered: *mean* (or expected value), *variance*, *median*, as well as the *first* and *third quartiles*. Except for mean and variance, which are of relatively high importance, we employ three percentiles which express the value below which a certain percent of observations may be found. The first, second (median), and third quartiles correspond to 25, 50, and 75 percent, respectively. The function which describes the short-term integration is the following:

$$\begin{aligned} F_{stat}(x_t, \ldots, x_{t+p-1}) = [&mean(x_t, \ldots, x_{t+p-1}), \\ &\mathrm{var}(x_t, \ldots, x_{t+p-1}), \ldots q1(x_t, \ldots, x_{t+p-1}), \\ &median(x_t, \ldots, x_{t+p-1}), q3(x_t, \ldots, x_{t+p-1})]. \end{aligned} \tag{2}$$

The dimension of the integrated vector is 5 times the original one; thus $R = 5\mathrm{x}D$. Obviously, this kind of approach does not capture the dynamicity that characterizes an audio signal since another combination of subsequent observations can be transformed to the same integrated vector. However, the next two approaches have the ability to capture the temporal behavior of a given series.

3.2 Spectral Moments

The spectral characteristics of a specific texture window may provide essential information with respect to relationships between successive observations. The STFT of a texture window forms the basis for calculating the spectral moments. With this procedure we can determine the sinusoidal frequency and phase content of local sections of a given sequence as it changes over time. It should be noted that another parameter is inserted here, the FFT size, which symbolizes the maximum number of frames which can be integrated (this parameter is irrelevant to the respective size of the feature extraction algorithms).

We calculate the power spectrum in dB of the series of a particular descriptor and store its mean value μ. Subsequently the next four statistics over the texture size of the amplitude spectrum are calculated: the mean m, variance v, skewness γ, and kurtosis κ. Thus, the final integrated vector has five times larger dimension than the starting one, like the previous temporal integration methodology ($R = 5\mathrm{x}D$). The formula for the spectral moments type of integration is the following:

$$F_{spec}(x_t, \ldots, x_{t+p-1}) = [\mu, m, v, \gamma, \kappa]. \tag{3}$$

3.3 Autoregressive Models

The logic behind the particular method is to model the temporal evolution of a specific feature sequence by trying to fit an AR process to it. The algorithms that were used are based on a stepwise least square approximation, which is computationally efficient when we are dealing with high-dimensional data [21]. Furthermore, confidence intervals for the estimated model parameters can be inferred for measuring how well the fitted model corresponds to the given data. In order to compute the integrated feature vector we make use of the respective coefficients of the produced autoregressive process. The coefficients of a joint autoregressive model A_n of order O can be inferred from the following equation:

$$x[t] = w + \sum_{n=1}^{O} x[t-n]A_n + e_t, \tag{4}$$

where w is the intercept vector, A_{n} are the $D\mathrm{x}D$ coefficient matrices of the autoregressive model, and e_{t} is a white noise vector of dimension D. Thus, the integrated feature vector for the multivariate autoregressive (MAR) modeling is given by the next formula:

$$F_{MAR}(x_t, \ldots, x_{t+p-1}) = [\vec{w}, \vec{A}_1, \ldots, \vec{A}_o], \tag{5}$$

which is of dimension $R = D(O\mathrm{x}D + 1)$. The same least-square approximations are computed for the case of diagonal autoregressive (DAR) modeling, but a further assumption is made: that the sound descriptors are independent with respect to each other. As a result the constraint exists that the model coefficients should be diagonal matrices. Hence, we calculate the parameters for each feature alone each time and the outcomes are concatenated. In this case we have a vector of significantly lower dimension, $R = D(O + 1)$:

$$F_{DAR}(x_t, \ldots, x_{t+p-1}) = [\vec{w}, \vec{D}_1, \ldots, \vec{D}_o]. \tag{6}$$

Finally, we employ the centered autoregressive (CAR) process too. The respective formula is the next one:

$$F_{CAR}(x_t, \ldots, x_{t+p-1}) = [\mu_t, \vec{D}_1, \ldots, \vec{D}_o], \tag{7}$$

where the respective coefficients are computed by the normalized (centered) process using the Levinson-Durbin algorithm [22]. In the specific case, the mean is calculated before the approximation of the AR process coefficients. Thus, the integration functions of DAR and CAR are different mainly because the CAR coefficients are computed on the centered feature vectors. During our experiments we employed only the latter two methods (DAR and CAR) due to the huge computational load inserted by the dimensionality of the vector which is produced by the MAR method.

4 Emotion Recognition Using Temporal Feature Integration

In this section we describe the experiments which were conducted in order to evaluate the efficiency of the methodologies analyzed so far. Our main aim is to assess the performance of the wavelet-based feature set as well as the efficacy of each temporal feature integration method. For each classification phase we employed left-right HMMs which were optimized in terms of number of states and Gaussian components. Furthermore, we used the Berlin Emotional Speech Database [23] and the leave-one-speaker-out (LOSO) method since we aim for a speaker-independent emotion recognition system. Here, we provide a brief analysis of the specific database for the sake of completeness. The "big six" emotion set as defined by the MPEG-4 protocol is represented in the Berlin Emotional Speech Database. More specifically, the following emotions are included: *anger*, *disgust*, *fear*, *joy*, *sadness*, and *surprise*. The corpus contains speech signals by 10 German professional actors, five of which are female. It should be noted that the sampling rate is 16 kHz with 16 bit analysis and the human perception test results in 84.3 percent average recognition rate.

4.1 Framework Parameterization

The window for the low level feature extraction algorithms is 30 ms with 20 ms overlap in order to ensure stability in case of

TABLE 2
The Recognition Rates with Respect to Each Group of Audio Parameters and Integration Methodology

Feature set	Integration methodology (order *O*)	Texture window (frames)	No. of states	No. of Gaussian components	Recall rate (%)	Precision rate (%)	F-measure (%)
Baseline set (Mel filter-bank, pitch, HNR)	no integration	-	3	128	68.4	66.3	67.3
	CAR (1)	50	6	16	66.25	67.5	66.9
	CAR (2)	**120**	**4**	**4**	**68.8**	**69.8**	**69.3**
	CAR (3)	90	4	8	65.3	64.3	64.8
	DAR (1)	120	4	32	55.5	57.2	56.3
	DAR (2)	120	5	16	57.9	56.7	57.3
	DAR (3)	90	4	8	55.9	56.6	56.2
	spectral moments	120	3	16	63.7	64.3	63.9
	short-term statistics	90	5	32	59.3	60.1	59.7
PWP integration analysis	no integration	-	3	8	63.2	62.9	63
	CAR (1)	30	3	4	61.9	65	63.4
	CAR (2)	30	4	4	60.5	54.4	57.3
	DAR (1)	30	3	2	43.2	47.7	45.3
	DAR (2)	120	3	16	40.8	43.3	42
	spectral moments	**20**	**4**	**8**	**63.5**	**67.5**	**65.4**
	short-term statistics	40	3	4	60.3	63.9	62

misalignments. We employ the Hamming windowing technique to smooth any possible discontinuities and the FFT size is 512. With respect to the PCA coefficients we selected the smallest number of components, which keeps at least 95 percent of the variance. By running an experiment on the training data for both groups of descriptors we arrived at the following results: 15 components for the Mel-filterbank-based set and 70 for the Perceptual Wavelet Packet integration analysis set. During each one of the experiments a PCA kernel was computed using the training data and then employed to transform the novel data to the training data-dependent coordination system.

We used the ARFIT toolbox [21] as a basis for implementing the temporal integration techniques which are based on autoregressive modeling (DAR and CAR). The ARFIT toolbox is written in MATLAB and includes functions for analyzing time series of multiple variables using AR processes. The FFT size of the spectral moments technique was set equal to 128; thus the system can integrate up to 128 subsequent frames which corresponds to about 2.5 seconds. The values of the frames that are to be integrated into a texture window were taken from the set: {10, 20, 30, 40, 50, 60, 90, 120}, while a constant hop-size of 10 frames was adopted so that the final number of texture windows was kept the same independently of the included number of frames. The particular numbers of frames were chosen after considering the next two parameters: 1) They comprise quite a large set, which enables a thorough test of temporal feature integration, and 2) the largest duration ($\sim$ 2.5 s) provides an output which is still beneficial to an HCI system which employs an emotion recognition framework, e.g., for adjusting its operation. It should be noted that if the sound sample is of small duration, then it is integrated into one texture window only. With respect to the HMMs, the range of the number of states was between 3 and 7, while the numbers of Gaussian components tested, respectively, were: {2, 4, 8, 16, 32, 64, 128}. The final values of these parameters were chosen using the highest recognition rate criterion.

4.2 Classification Results

This section provides the classification results with respect to the different levels of our study. Initially, we compare the results of the different feature sets which were employed throughout the current study along with the temporal integration approaches. Subsequently, we provide a useful diagram which demonstrates the effect of the length of the texture window on the average classification rate. Finally, we draw our conclusions with respect to the feature set and integration technique with the highest recognition accuracy.

We employed recall rate, precision rate, and the f-measure in order to evaluate all the approaches:

$$recall = \frac{\#(correctly\_recognized\_recordings)}{\#(correctly\_recognized\_rec.) + \#(misclassified\_rec.)}, \tag{8}$$

$$precision = \frac{\#(correctly\_recognized\_recordings)}{\#(all\_recognized\_recordings)}, \tag{9}$$

$$f\text{-}measure = \frac{2^{*}recall^{*}\ precision}{recall + precision}. \tag{10}$$

The classification results are tabulated in Table 2. We can also observe the classification results without any type of temporal integration, which allows us to achieve a thorough comparison. Furthermore, the parameters of the HMMs (number of states and Gaussian components) are given for each case. It should be noted that the rates were averaged across all speakers. As we can see, the best recall rate is provided by the baseline set, while it is integrated according to the centered autoregressive technique of second order. The best results with respect to the set derived from the wavelet

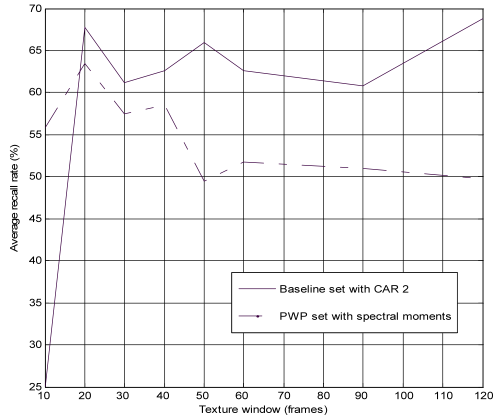

Fig. 3. Average recognition rate as a function of the length of the texture window. Each feature set is integrated according to the technique which provides the highest recognition accuracy.

domain are obtained using the spectral moments strategy for integration. However, the recognition gain is relatively small (0.4 percent for the baseline set and 0.3 percent for the PWP set) when compared to the nonintegrated feature set. A general observation is that most integration approaches produce lower average recall rates than the ones output by the system which employs the nonintegrated feature set. This reveals that the integrated vector does not always capture aspects of a specific emotion class, which are essential for automatic recognition. However, it is clear that the DAR integration method produces lower f-measures when compared to the CAR method. This fact demonstrates the efficacy on the recognition performance of the Levinson-Durbin algorithm [22] which is employed on the a priori centered feature vectors.

An interesting observation is that the Mel filterbank-based set provides the highest recall rates when the order of the autoregressive processes is equal to 2. This fact does not hold for the PWP set since its recognition rates drop as the order of the autoregressive processes increases. Thus, in the case of the PWP set, we did not run experiments with $O = 3$.

One can make the logical assumption that the larger the value of the integration window, the higher the recognition rates will be since more information can be exploited. However, this assumption is not true in every case. During this phase of the experimental results analysis we isolated the best recall rate for each texture window with respect to each group of parameters according to the integration method which provided the highest recall rate. It should be noted that the number of the training texture windows that were employed for HMM construction was constant due to the fixed value of the hop size; hence we avoided overfitting because of insufficient data. Fig. 3 illustrates the variation that the average recognition rate exhibits as the length of the texture window changes. As we can see, the set based on the multiresolution analysis presents a peak when the texture window length is equal to 20 and then it follows a decreasing course. The baseline set demonstrates better performance across almost all texture window lengths, while its highest rate is presented when we integrate 120 frames. We infer that temporal integration is useful only to some extent which heavily depends on the specific set of acoustic parameters. After a certain limit, the integrated information does not exhibit a consistent pattern, which hinders the model construction procedure.

During this experimental phase we identified that there are some classes of emotions which are classified more precisely by the baseline set or the PWP set, which indicates a degree of complementarity among these two feature sets. They share some common characteristics, but also exhibit many differences when it comes to discriminate particular emotions. More specifically, the emotional states of fear and disgust are recognized more accurately when using the wavelet-based feature set. The recognition rate which is provided by the baseline feature set surpasses the wavelet-based one when it comes to happiness and anger. With regard to the last two emotional states (neutral and sadness), we did not observe significant differences between the respective recognition accuracies. Hence, we decided to explore the usage of several fusion techniques which elaborate on various levels. This experimental phase is described in the next section.

5 Multilevel Fusion for Emotion Recognition

Following the observations of the previous section we experimented on fusing information on the next levels: 1) feature level fusion, 2) fusion of the log likelihoods which are produced by the temporally integrated feature sets, and 3) fusion of temporal integration methods. The latter two methodologies belong to the integration type, which is often called *late integration*. The corresponding results are organized in Table 3, where the highest average recognition rates for each fusion type are highlighted. With respect to the feature level fusion we employed the HMM approach, while our experimentations included both temporally integrated sets and not. Subsequently, we fused the outputs of the HMMs which were created using both integrated and nonintegrated feature sequences. The last phase of the fusion experiments concerned the exploitation of the three different temporal integration methods concurrently. During the specific phase we used each feature set independently.

The LOSO approach was adopted during the series of experiments which included fusion as well. We evaluated the performance of five fusion schemes: multilayer perceptron, J48 tree, random forest, support vector classifier (SMO), and the simple logistic method [24]. Here, we are going to provide only a brief analysis of these methods, as a complete description lies outside the scope of the current work. Decision trees can be easily constructed in a supervised way while no assumption is made a priori about the distribution of the data. Their main disadvantage is that slight alterations of the training set can result in a decision tree structure that exhibits a large number of differences. However, we reduce this risk since we elaborate on probabilities of feature sequences and not on the features themselves. The multilayer perceptron method follows the logic of the linear perceptron, while it employs nodes with nonlinear activation functions for discriminating data that are not linearly separable. Additionally, artificial neural networks can be very useful where the patterns are not evident. The backpropagation algorithm was used to train the neural network with one hidden layer.

TABLE 3
The Recognition Rates Obtained by Using Fusion at Different Levels

Fusion level	Pattern recognition algorithm	Recall rate (%)	Precision rate (%)	F-measure (%)
Feature without temporal integration	HMM (2 modes, 4 states)	63.2	64.4	63.8
Feature with temporal integration	HMM (8 modes, 3 states)	64.5	66.7	65.6
Log-likelihood without temporal integration	**MLP**	**65.6**	**69.5**	**67.5**
	J.48 Tree	52.2	54.4	53.3
	Random forest	57.7	58	57.8
	SMO	39.5	43.3	41.3
	Simple logistic	54.2	53.7	53.9
Log-likelihood with optimally integrated feature sets	MLP	90.7	88.6	89.6
	J.48 Tree	67.5	71.4	69.4
	Random forest	70.2	69.4	69.8
	SMO	91.1	90.5	90.8
	Simple logistic	**92.2**	**93.4**	**92.8**
Log-likelihood of temporal integration methods – *Baseline* feature set	MLP	61.3	62.3	61.8
	J.48 Tree	84.7	83.9	84.3
	Random forest	90.9	91.3	91.1
	SMO	91.3	**92.4**	**91.8**
	Simple logistic	**91.6**	91.9	91.7
Log-likelihood of temporal integration methods – *PWP* feature set	**MLP**	**91.8**	92.3	**92**
	J.48 Tree	90.8	93.3	**92**
	Random forest	82.1	87.5	84.7
	SMO	87.8	85.6	86.7
	Simple logistic	90.2	**93.6**	91.9

The number of its nodes was equal to half the total of the number of features plus the number of classes at a learning rate of 0.3. The rest of the classification schemas were applied using their default parameters. The Sequential Minimal Optimization classifier is trained in such a manner that the decision function maximizes the generalization ability of the classifier. The latter is achieved by finding the optimal separating hyperplane that maximizes the margin between the support vectors of the six classes of emotions. Here, the term *support vectors* refers to these class-specific representatives of training data that are important cornerstones in the estimation of the optimal separating hyperplane. Last, the linear logistic regression modeling is based on implementing ordinal regression functions as base learners to fit the logistic models, while cross validation is employed to acquire the optimal number of iterations that offer automatic attribute selection [25]. It should be noted that the log likelihoods produced were normalized by the number of the integrated texture windows.

These methods were chosen because of their ability to handle redundant data, which means that in the case where the feature sets capture overlapping information of the acoustic signal that they represent, the algorithm can effectively exploit it and the performance of the recognizer is usually increased when compared to employing each feature set alone. A redundant feature set may provide improved performance under adverse conditions (where parts of the signal's spectrum are distorted or simply missing) as is the case in real life.

As we can observe in Table 3, both types of feature-level fusion did not provide satisfactory results. In fact, their performance is lower than using one feature set alone while the computational complexity is increased due to the fusion modeling process. Similarly, the fusion at the log-likelihood level without temporal integration provided 65.6 percent average recognition rate using the MLP methodology, in spite of the additional complexity which is inserted by the temporal integration methodologies. Afterward, we elaborated on fusing the log likelihoods which are produced by the models constructed using the optimally integrated feature sets, that is, second order centered autoregressive process for the baseline set and spectral moments for the PWP set (see Section 4.2). The particular experiment produced very good results, reaching 92.2 percent when the simple logistic regression method is employed. Finally, we experimented on fusing the information as it is modeled by different temporal integration techniques since they may capture different and distinctive characteristics of the audio structure. More precisely, we isolated the models which produced the highest average recognition rates with respect to each feature set and integration technique (these models can be inferred from Table 2). The results are quite promising, as the Mel-filterbank-based and the PWP sets produced 91.6 and 91.8 percent average recall rates, respectively. It is worth noting that during the particular

TABLE 4
Confusion Matrix of the System
Which Is Based on Log-Likelihood Level Fusion
of Optimally Integrated Sets of Acoustic Parameters

Responded Presented	Anger	Happy	Neutral	Disgust	Fear	Sadness
Anger	**96,4**	3,1	0,5	0	0	0
Happy	2,1	**88,8**	7,7	0	1,4	0
Neutral	0	0	**95,8**	2,5	1,7	0
Disgust	0	0	9,1	**90,9**	0	0
Fear	7	1,1	0	0	**91,9**	0
Sadness	0	0	0	0	10,6	**89,4**

experimental stage the PWP set surpassed slightly the baseline one. Furthermore, the baseline set presents its best performance when using the logistic regression method, while the highest recognition rate of the PWP set is exhibited when using the MLP fusion scheme. The log likelihoods produced by the HMMs with temporally integrated features are more suitable for fusion than the log likelihoods produced by the HMMs with no temporally integrated features, since the integrated feature sequence is more likely to exhibit a more consistent pattern. Thus, the training algorithm which elaborates on these data may achieve finer modeling.

The confusion matrix of the methodology which produced the highest rate is shown in Table 4. As we can see, the system achieves high recognition rates across all emotions. The best rate is equal to 96.4 percent and corresponds to the rate achieved for recognizing the emotional states of anger. The lowest rate is 88.8 percent, which corresponds to the ability of the system with regard to classifying the audio signals which contain expressions of the happy emotional state. The most serious misclassifications concern the happy class, which is confused with the neutral one, and the fear class, which is confused with manifestations of anger. However, these misclassifications are expected due to the similarity of the respective audio data. Furthermore, many of the misclassifications occur due to the great variability among sound samples of the same class as can be inferred by a human listener.

We infer that the performance of the system is excellent when it is based on the concurrent usage of diverse groups of sound parameters which are optimally integrated (in the temporal sense). This suggests that the problem of speech emotion recognition is better handled when using a multidomain group of descriptors. Moreover, our methodology significantly outperforms the ability of a human listener to classify the respective signals which is equal to 84.3 percent. We conclude that the final classification results are more than encouraging and demonstrate the applicability of the proposed method on the specific research domain.

6 Conclusion

This paper presented a framework for speech emotion recognition which is based on feature sets from diverse domains as well as on modeling their evolution in time. The experimental protocol was carefully designed and all the aspects of the suggested methodology were evaluated in detail. It was shown that it leads to high accuracy with respect to recognizing six emotional states. We demonstrated the merits of temporal feature integration as well as fusion on the log-likelihood level. Moreover, the results reveal that different texture windows are appropriate for each feature set along with a different temporal integration algorithm. Additional classes of various emotions which are currently not included in our work can be easily incorporated as long as a sufficient amount of training data is collected. Our general conclusion is that the problem of automated recognition of human emotions via speech is better addressed while utilizing multidomain groups of descriptors. Combined, they provide information which is essential for automatic emotion recognition. A slight disadvantage is the fact that the system needs more time to make a prediction due to the increased need of information, but this is not critical for most applications.

The proposed framework is flexible and can facilitate many audio recognition applications. Our future steps include the collection of real-world data and the application of the current approach. This is toward evaluating the specific method on real-world HCI applications, such as an automated smart-home assistant, which is characterized by a high degree of difficulty since we are dealing with naturalistic data. In the specific case, we are going to analyze the limitations of the proposed method which models the temporal evolution of acoustic parameters. When dealing with naturalistic data the acoustic patterns might present additional difficulties during training and recognition. It is usual that some frequency bands are absent of distorted (for example, by environmental noise). These difficulties could be addressed by the insertion of additional features (which better capture acoustic characteristics linked to the specific problem) and/or by employing multiple classifier systems [2].

Acknowledgments

This work was supported by the EC FP 7th grant PLAYMANCER 215839 "A European Serious Gaming 3D Environment."

References

[1] C.M. Lee and S.S. Narayanan, "Toward Detecting Emotions in Spoken Dialogs," *IEEE Trans. Speech and Audio Processing,* vol. 13, no. 2, pp. 293-303, Mar. 2005.

[2] M.E. Ayadi, M.S. Kamel, and F. Karray, "Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases," *Pattern Recognition,* vol. 44, no. 3, pp. 572-587, Mar. 2011.

[3] G. Zhou, J.H.L. Hansen, and J.F. Kaiser, "Nonlinear Feature Based Classification of Speech under Stress," *IEEE Trans. Speech and Audio Processing,* vol. 9, no. 2, pp. 201-216, Mar. 2001.

[4] Y. Li and Y. Zhao, "Recognizing Emotions in Speech Using Short-Term and Long-Term Features," *Proc. Int'l Conf. Spoken Language Processing,* pp. 2255-2258, 1998.

[5] D.N. Jiang and L.-H. Cai, "Speech Emotion Classification with the Combination of Statistic Features and Temporal Features," *Proc. Int'l Conf. Multimedia and Expo,* pp. 1967-1970, 2004.

[6] B. Vlasenko, B. Schuller, A. Wendemuth, and G. Rigoll, "Combining Frame and Turn-Level Information for Robust Recognition of Emotions within Speech," *Proc. Int'l Conf. Spoken Language Processing,* pp. 2225-2228, 2007.

[7] B. Vlasenko, B. Schuller, A. Wendemuth, and G. Rigoll, "Frame vs. Turn-Level: Emotion Recognition from Speech Considering Static and Dynamic Processing," *Proc. Second Int'l Conf. Affective Computing and Intelligent Interaction,* A. Paiva, ed., pp. 139-147, 2007.

[8] S. Wu, T.H. Falk, and W.-Y. Chan, "Automatic Recognition of Speech Emotion Using Long-Term Spectro-temporal Features," *Proc. Int'l Conf. Digital Signal Processing,* pp. 205-210, 2009.

[9] A. Meng, P. Ahrendt, J. Larsen, and L.K. Hansen, "Temporal Feature Integration for Music Genre Classification," *IEEE Trans. Audio, Speech, and Language Processing,* vol. 15, no. 5, pp. 1654-1664, July 2007.

[10] C. Joder, S. Essid, and G. Richard, "Temporal Integration for Audio Classification with Application to Musical Instrument Classification," *IEEE Trans. Audio, Speech, and Language Processing,* vol. 17, no. 1, pp. 174-186, Jan. 2009.

[11] S. Ntalampiras, I. Potamitis, and N. Fakotakis, "Exploiting Temporal Feature Integration for Generalized Sound Recognition," *EURASIP J. Advances in Signal Processing,* vol. 2009, Article ID 807162, 2009, doi:10.1155/2009/807162.

[12] G. Zhou, J.H.L. Hansen, and J.F. Kaiser, "Methods for Stress Classification: Nonlinear Teo and Linear Speech Based Features," *Proc. IEEE Int'l Conf. Acoustics and Signal Processing,* pp. 2087-2090, 1999.

[13] R. Fernandez and R.W. Picard, "Modeling Drivers' Speech under Stress," *Proc. Int'l Speech Comm. Assoc. Workshop Speech and Emotions,* 2000.

[14] S. Ntalampiras, I. Potamitis, and N. Fakotakis, "A Multidomain Approach for Automatic Home Environmental Sound Classification," *Proc. 11th Ann. Conf. the Int'l Speech Comm.,* pp. 2210-2213, 2010.

[15] B. Scharf, *Critical Bands, in Foundations of Modern Auditory Theory,* J.V. Tobias, ed., pp. 157-202. Academic Press, 1970.

[16] W.A. Yost, *Fundamentals of Hearing,* third ed., pp. 153-167. Academic Press, 1994.

[17] Torch Machine Learning Library, http://www.torch.ch, 2012.

[18] J.-J. Aucouturier, B. Defreville, and F. Pachet, "The Bag-of-Frames Approach to Audio Pattern Recognition: A Sufficient Model for Urban Soundscapes but Not for Polyphonic Music," *J. Acoustical Soc. of Am.,* vol. 122, no. 2, pp. 881-891, Aug. 2007.

[19] M.M.H. El Ayadi, M.S. Kamel, and F. Karray, "Speech Emotion Recognition Using Gaussian Mixture Vector Autoregressive Models," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing,* vol. 4, pp. 957-960, 2007.

[20] L. Fu, X. Mao, and L. Chen, "Speaker Independent Emotion Recognition Based on SVM/HMMs Fusion System," *Proc. Int'l Conf. Audio, Language, and Image Processing,* pp. 61-65, 2008.

[21] T. Schneider and A. Neumaier, "Algorithm 808: ARFIT—A Matlab Package for the Estimation of Parameters and Eigenmodes of Multivariate Autoregressive Models," *ACM Trans. Math. Software,* vol. 27, no. 1, pp. 58-65, Mar. 2001.

[22] P. Delsarte and Y.V. Genin, "The Split Levinson Algorithm," *IEEE Trans. Acoustics, Speech, and Signal Processing,* vol. 34, no. 3, pp. 470-478, June 1986.

[23] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A Database of German Emotional Speech," *Proc. Int'l Conf. Spoken Language Processing,* pp. 1517-1520, 2005.

[24] I.H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques,* second ed. Morgan Kaufmann, 2005.

[25] N. Landwehr, M. Hall, and E. Frank, "Logistic Model Trees," *Proc. European Conf. Machine Learning,* pp. 241-252, 2003.

Stavros Ntalampiras received the engineer's and PhD degrees from the Department of Electrical and Computer Engineering of the University of Patras, Greece, in 2006 and 2010, respectively. Since 2010 he has been a postdoctoral researcher in the Wire Communications Laboratory of the same department. His research interests are content-based signal processing, audio pattern recognition, and computer audition. He has authored more than 20 publications in refereed journals and conferences.

Nikos Fakotakis received the BSc degree from the University of London, United Kingdom, in electronics in 1978, the MSc degree in electronics from the University of Wales, Cardiff, United Kingdom, in 1979, and the PhD degree in speech processing from the University of Patras, Greece, in 1986. From 1986 to 1992 he was a lecturer in the Electrical and Computer Engineering Department of the University of Patras, from 1992 to 1999 an assistant professor, from 2000 to 2003 an associate professor, and since 2003, he has been a full professor in the area of speech and natural language processing. He is a director of the Communication and Information Technology Division, director of the Wire Communications Laboratory (WCL), and head of the Artificial Intelligence Group. He is an author or coauthor of more than 300 publications in the areas of speech and natural language engineering and artificial intelligence. His current research interests include AI, speech recognition/understanding, speaker recognition, user modeling, spoken dialogue processing, and natural language processing.