D. Ververidis and C. Kotropoulos, "Fast Sequential Floating Forward Selection applied to emotional speech features estimated on DES and SUSAS data collections," in Proc. European Signal Processing Conf. (EUSIPCO), Italy, 2006.

FAST SEQUENTIAL FLOATING FORWARD SELECTION APPLIED TO EMOTIONAL SPEECH FEATURES ESTIMATED ON DES AND SUSAS DATA COLLECTIONS

Dimitrios Ververidis and Constantine Kotropoulos

Department of Informatics, Aristotle University of Thessaloniki
Box 451, Thessaloniki 541 24, Greece
E-mail: {jimver, costas}@zeus.csd.auth.gr

ABSTRACT

In this paper, we classify speech into several emotional states based on the statistical properties of prosody features estimated on utterances extracted from Danish Emotional Speech (DES) and a subset of Speech Under Simulated and Actual Stress (SUSAS) data collections. The proposed novelties are in: 1) speeding up the sequential floating feature selection up to 60%, 2) applying fusion of decisions taken on short speech segments in order to derive a unique decision for longer utterances, and 3) demonstrating that gender and accent information reduce the classification error. Indeed, a lower classification error by 1% to 11% is achieved, when the combination of decisions is made on long phrases and an error reduction by 2%-11% is obtained, when the gender and the accent information is exploited. The total classification error reported on DES is 42.8%. The same figure on SUSAS is 46.3%. The reported human errors have been 32.3% in DES and 42% in SUSAS. For comparison purposes, a random classification would yield an error of 80% in DES and 87.5% in SUSAS, respectively.

1. INTRODUCTION

Emotional speech classification is a problem that has attracted recently the interest of scientific community [1, 2]. In this paper, the sequential floating forward selection algorithm is used for feature selection in order to minimize the emotion classification error of the Bayes classifier when the class conditional probability distribution functions (pdfs) of features are modeled as Gaussians. To estimate the classification error achieved by the Bayes classifier, crossvalidation is employed [3]. A technique is proposed that guarantees statistically significant reductions of the classification error committed by the Bayes classifier, when new features are added. The aforementioned technique controls the number of crossvalidation repetitions in sequential forward feature selection algorithms. Frequently, the emotional speech classification is conducted on utterances, i.e. speech segments between two silence pauses. However, the human evaluators provide ground truth for phrases that consist of sentences and paragraphs. The median rule for decision fusion is proposed in order to combine the decisions taken by processing utterances separately and to derive a unique decision for phrases.

The outline of the paper is as follows. In Section 2, the speech utterances extracted from the data collections employed and the prosody features extracted from the speech utterances are described. Section 3 is devoted to the estimation of the classification error committed by the Bayes classifier during crossvalidation repetitions when the class conditional pdfs of the prosody features are modeled by Gaussians. A mechanism that controls the number of crossvalidation repetitions is developed in the next section. This mechanism is incorporated into the sequential floating forward selection algorithm to speed up its execution. In Section 5, we propose an algorithm to fuse decisions taken on short speech segments in order to derive a unique decision for long phrases and to reduce the classification error. Experimental results on speeding up feature selection, fusing decisions, and exploiting accent and gender information are demonstrated in Section 6. Finally, conclusions are drawn in Section 7.

This work has been supported by the research project 01ED312 "Use of Virtual Reality for training pupils to deal with earthquakes" financed by the Greek Secretariat of Research and Technology.

2. DATA AND FEATURE EXTRACTION

Two data collections specific to emotion recognition are exploited. The first data collection is the Danish Emotion Speech (DES) [4] whose recordings refer to speech expressed by 2 male and 2 female actors in 5 emotional states such as *anger, happiness, neutral, sadness*, and *surprise*. The speech data consist of 2 isolated words, 9 isolated sentences, and 2 isolated paragraphs. *Set* $\mathcal{A}$ is formed by 360 utterances corresponding to words and sentences. *Set* $\mathcal{B}$ is the union of *Set* $\mathcal{A}$ and another 800 utterances extracted from paragraphs. In the experiments, Set $\mathcal{A}$ and Set $\mathcal{B}$ are divided into subsets $\mathcal{A}_m$, $\mathcal{A}_f$ and subsets $\mathcal{B}_m$, $\mathcal{B}_f$ for male and female speakers, respectively. The second data collection uses a part of the Speech Under Simulated and Actual Stress (SUSAS) data collection [5] and is denoted as Set $\mathcal{C}$. $\mathcal{C}$ includes speech utterances under low and high stress conditions (the so-called *Cond50* and *Cond70*, respectively) and speech under various talking styles such as *anger, clear, fast, loud, question, slow*, and *soft*. Data from 9 male speakers with three regional accents, i.e. that of Boston, General, and New York are exploited. Set $\mathcal{C}$ is divided into subsets $\mathcal{C}_B$, $\mathcal{C}_G$, and $\mathcal{C}_N$ corresponding to the aforementioned three regional accents.

The so-called global statistics of prosody feature contours [6], i.e., statistical properties of *pitch, formants*, and *energy* features are used. The prosody features are estimated on a frame basis, $f_s(n;m) = s(n)w(m-n)$, where $s(n)$ is the speech signal and $w(m-n)$ is a window of length N_w ending at sample m [7]. The trends of the feature contours (i.e. *plateaux* at *minimum/maxima* or *rising/falling slopes*) is a valuable feature for emotion recognition because they describe the temporal characteristics of emotions. In the following, the methods to extract pitch, formants, and energy features, as well as the technique to track contour slopes and plateaux are described.

The *pitch signal*, also known as glottal waveform, has information about emotion, because it depends on the tension of the vocal folds and the subglottal air pressure. The pitch signal is produced from the vibration of the vocal folds. The time elapsed between two successive vocal fold openings is called *pitch period* T, while the vibration rate of the vocal folds is the *fundamental frequency of the phonation* F_0 or *pitch frequency*. The method used for extracting pitch is based on the *autocorrelation of center-clipped* frames. The signal is low filtered at 900 Hz and then it is segmented to short-time frames of speech $f_s(n;m)$. The clipping, which is a non-linear procedure that prevents the 1st formant interfering with the pitch, is applied to each frame $f_s(n;m)$ yielding

$$\hat{f}_s(n;m) = \begin{cases} f_s(n;m) - \Lambda & \text{if } |f_s(n;m)| > \Lambda \\ 0 & \text{if } |f_s(n;m)| < \Lambda \end{cases} \quad \forall n \qquad (1)$$

where Λ is set at the 30% of the maximum value of $f_s(n;m)$. The

pitch frequency is estimated by the short-term autocorrelation

$$r_s(\lambda; m) = \frac{1}{N_w} \sum_{n=m-N_w+1}^{m} \hat{f}_s(n; m)\hat{f}_s(n-\lambda; m) \quad (2)$$

where λ is the lag. The pitch frequency of the frame ending at m is given by

$$\hat{F}_0(m) = \frac{F_s}{N_w} \operatorname{argmax}_\lambda \{|r_s(\lambda; m)|\}_{\lambda=N_w\,(F_l/F_s)}^{\lambda=N_w\,(F_h/F_s)} \quad (3)$$

where F_s is the sampling frequency, and F_l, F_h are the perceived lowest and highest possible pitch frequencies by humans, respectively. The values of the aforementioned parameters are $F_s = 8000$ Hz, $F_l = 50$ Hz, and $F_h = 300$ Hz.

The method to estimate *formants* relies on the *linear prediction analysis*. Let a 10-order all-pole vocal tract model at frame m $\hat{\Theta}(z; m)$ with *linear prediction coefficients* (LPCs) $\hat{a}_\zeta(m)$ be

$$\hat{\Theta}(z; m) = \frac{1}{1 - \sum_{\zeta=1}^{10} \hat{a}_\zeta(m) z^{-\zeta}} = \frac{1}{\prod_{\zeta=1}^{10} (z - p_\zeta(m))}. \quad (4)$$

In (4), $\hat{a}_\zeta(m)$ are estimated by the Levinson-Durbin algorithm and the order of the model for speech sampled at 8 kHz is selected as 10. The angles of the 4 poles $p_\zeta(m)$ which are furthest from the origin are indicators of the 4 formant frequencies. The energy of the speech frame ending at m is

$$e(m) = \frac{1}{N_w} \sum_{n=m-N_w+1}^{m} |f_s(n; m)|^2. \quad (5)$$

In order to find the energy content of a frequency band, a FIR filter of 120 coefficients is employed. The coefficients are calculated with the frequency sampling method using a Hamming window.

A *contour* of a short-term feature is formed by assigning the feature value computed on a frame basis to all samples belonging to the frame. For example, the energy contour is given by

$$E(n) = e(m), \quad n = m - N_w + 1, \ldots, m. \quad (6)$$

The contour $E(n),\ n = 1, 2, \ldots, L$, where L is the length of the signal, is smoothed by applying a moving average operator of 100 data points, resulting to $\hat{E}(n)$. To determine which samples belong to a set of rising slopes ($\mathcal{S}_r$), falling slopes ($\mathcal{S}_f$), plateaux at maxima ($\mathcal{S}_{ma}$), and plateaux at minima ($\mathcal{S}_{mi}$), the first derivative of the feature contour is estimated by numerical methods. The derivative of the energy contour is estimated by the first-order difference $\hat{E}_D(n) = \hat{E}(n) - \hat{E}(n-1)$, $n = 2, \ldots, L$. Subsequently, the algorithm of Figure 1 is applied. In this algorithm, $v_1 = 10^{-3}$ is a constant that enables the detection of the rising or falling slopes and the plateaux. The distinction between the plateaux at maxima and those at minima is accomplished with the constant v_2 which is set to 0.45. The statistical features employed in this study are grouped

---

if $\hat{E}_D(n) \geq v_1$, $s(n) \in \mathcal{S}_r$
else if $\hat{E}_D(n) \leq -v_1$, $s(n) \in \mathcal{S}_f$
else if $|\hat{E}_D(n)| < v_1$
  if $E(n) > \max(E(i)) \cdot v_2$, $s(n) \in \mathcal{S}_{ma}$
  else if $E(n) \leq \max(E(i)) \cdot v_2$, $s(n) \in \mathcal{S}_{mi}$
  end
end

Figure 1: Algorithm for finding the plateaux at minima/maxima and the rising/falling slopes of pitch and energy contours.

in several classes as is explained subsequently. The features are referenced by their corresponding indices throughout the analysis following.

2.1 Formants features

The set of formants features is comprised by the statistical properties of the 4 formant frequency contours.
1. - 4. Mean value of the first, second, third, and fourth formant
5. - 8. Maximum value of the first, second, third, and fourth formant
9. - 12. Minimum value of the first, second, third, and fourth formant
13. - 16. Variance of the first, second, third, and fourth formant

2.2 Pitch features

The pitch features are statistics of the pitch frequency contour.
17. - 21. Maximum, minimum, mean, median, interquartile range of pitch values
22. Pitch existence in the utterance expressed in percentage (0-100%)
23. - 26. Maximum, mean, median, interquartile range of durations for the plateaux at minima
27. - 29. Mean, median, interquartile range of pitch values for the plateaux at minima
30. - 34. Maximum, mean, median, interquartile range, upper limit (90%) of durations for the plateaux at maxima
35. - 37. Mean, median, interquartile range of the pitch values within the plateaux at maxima
38. - 41. Maximum, mean, median, interquartile range of durations of the rising slopes of pitch contours
42. - 44. Mean, median, interquartile range of the pitch values within the rising slopes of pitch contours
45. - 48. Maximum, mean, median, interquartile range of durations of the falling slopes of pitch contours
49. - 51. Mean, median, interquartile range of the pitch values within the falling slopes of pitch contours

2.3 Energy (intensity) features

The energy features are statistics of the energy contour.
52. - 56. Maximum, minimum, mean, median, interquartile range of energy values
57. - 60. Maximum, mean, median, interquartile range of durations for the plateaux at minima
61. - 63. Mean, median, interquartile range of energy values for the plateaux at minima
64. - 68. Maximum, mean, median, interquartile range, upper limit (90%) of duration for the plateaux at maxima
69. - 71. Mean, median, interquartile range of the energy values within the plateaux at maxima
72. - 75. Maximum, mean, median, interquartile range of durations of the rising slopes of energy contours
76. - 78. Mean, median, interquartile range of the energy values within the rising slopes of energy contours
79. - 82. Maximum, mean, median, interquartile range of durations of the falling slopes of energy contours
83. - 85. Mean, median, interquartile range of the energy values within the falling slopes of energy contours

2.4 Spectral features

The spectral features is the energy content of certain frequency bands divided to the length of the utterance.
86. - 93. Energy below 250, 600, 1000, 1500, 2100, 2800, 3500, and 3950 Hz.
94.-100. Energy in the 250 - 600, 600 - 1000, 1000 - 1500, 1500 - 2100, 2100 - 2800, 2800 - 3500, 3500 - 3950 frequency bands.
101.-106. Energy in the 250 - 1000, 600 - 1500, 1000 - 2100, 1500 -2800, 2100 - 3500, and 2800 - 3950 frequency bands.
107.-111. Energy in the 250 - 1500, 600 - 2100, 1000 - 2800, 1500 -3500, and 2100 - 3950 frequency bands.
112.-113. Energy ratio of (3950 - 2100)/(2100 - 0) and (2100-1000) /(1000 - 0) frequency bands.

To facilitate the classifier design, feature subset selection is needed. A criterion for comparing feature sets is as follows.

3. CROSSVALIDATION ERROR ESTIMATION

Let us denote the set of utterances by $\mathbf{u}^{\mathcal{W}} = \{\mathbf{u}_i^{\mathcal{W}}\}_{i=1}^{N}$. Such a set can be considered as an independent and identically distributed sample from the multidimensional distribution F of the feature set $\mathcal{W} = \{w_k\}_{k=1}^{K}$ which consists of K=113 features w_k. Each utterance $\mathbf{u}_i^{\mathcal{W}} = (\mathbf{y}_i^{\mathcal{W}}, l_i)$ is treated as a pattern consisting of a measurement vector $\mathbf{y}_i^{\mathcal{W}}$ and a label $l_i \in \{1, 2, \ldots, C\}$, where C is the total number of emotional states.

Let us predict the label of an utterance by processing the feature vectors using for example a classifier. A usual estimate of the prediction error using the sample $\mathbf{u}^{\mathcal{W}}$ is the cross-validation (CV) estimate. The CV estimate of prediction error is the mean of $b = \{1, 2, \ldots, B\}$ estimates of the error rate calculated as follows. In the bth repetition, $N_{\mathcal{D}} < N$ samples are randomly selected from $\mathbf{u}^{\mathcal{W}}$ without re-substitution to build the design set $\mathbf{u}_{\mathcal{D}b}^{\mathcal{W}}$, while the remaining set $\mathbf{u}_{\mathcal{T}b}^{\mathcal{W}}$ of $N_{\mathcal{T}} = N - N_{\mathcal{D}}$ samples creates the test set.

Let $Q[l_i, \eta_{\mathbf{u}_{\mathcal{D}b}^{\mathcal{W}}}(\mathbf{y}_i)]$ denote the zero-one loss function between the label l_i and its prediction for an utterance. For an utterance $\mathbf{u}_i^{\mathcal{W}} = (\mathbf{y}_i^{\mathcal{W}}, l_i)$, the prediction η is a discrete random variable admitting the value η if

$$\eta = \arg\max_{c=1}^{C} \{p_b(\mathbf{y}_i^{\mathcal{W}}|\Omega_c) P(\Omega_c)\}, \tag{7}$$

where $P(\Omega_c) = N_c/N$, N_c is the number of utterances that belong to class Ω_c with $c = \{1, 2, \ldots, C\}$, and $p_b(\mathbf{y}_i^{\mathcal{W}}|\Omega_c)$ is the class pdf of the measurement vector $\mathbf{y}_i^{\mathcal{W}}$ given Ω_c in the bth CV repetition. The class conditional pdf is assumed as a single Gaussian. Two parameters for each class Ω_c are required for a Gaussian, namely the mean vector $\boldsymbol{\mu}_c$ and the covariance matrix $\boldsymbol{\Sigma}_c$, $\forall \mathbf{y}_i^{\mathcal{W}} : \mathbf{u}_i^{\mathcal{W}} \in \Omega_c$. If $\mathbf{u}_{\mathcal{D}bc}^{\mathcal{W}} = \{\mathbf{u}_{\mathcal{D}b}^{\mathcal{W}} \cap \Omega_c\}$, then in a single CV repetition b the mean vector and the covariance matrix of each class Ω_c are

$$\boldsymbol{\mu}_{bc}^{\mathcal{W}} = \frac{1}{N_{\mathcal{D}}} \sum_{\mathbf{u}_i^{\mathcal{W}} \in \mathbf{u}_{\mathcal{D}bc}^{\mathcal{W}}} \mathbf{y}_i^{\mathcal{W}}, \tag{8}$$

$$\boldsymbol{\Sigma}_{bc}^{\mathcal{W}} = \frac{1}{N_{\mathcal{D}}} \sum_{\mathbf{u}_i^{\mathcal{W}} \in \mathbf{u}_{\mathcal{D}bc}^{\mathcal{W}}} (\mathbf{y}_i^{\mathcal{W}} - \boldsymbol{\mu}_{bc}^{\mathcal{W}})(\mathbf{y}_i^{\mathcal{W}} - \boldsymbol{\mu}_{bc}^{\mathcal{W}})^T. \tag{9}$$

The class conditional probability for each class Ω_c is

$$p_b(\mathbf{y}_i^{\mathcal{W}}|\Omega_c) = \frac{\exp[-\frac{1}{2}(\mathbf{y}_i^{\mathcal{W}} - \boldsymbol{\mu}_{bc}^{\mathcal{W}})^T (\boldsymbol{\Sigma}_{bc}^{\mathcal{W}})^{-1} (\mathbf{y}_i^{\mathcal{W}} - \boldsymbol{\mu}_{bc}^{\mathcal{W}})]}{(2\pi)^{K/2} |\det(\boldsymbol{\Sigma}_{bc}^{\mathcal{W}})|^{1/2}}, \tag{10}$$

where $\det(\cdot)$ is the determinant of a matrix. If $err(\hat{F}(\mathbf{u}_{\mathcal{D}b}^{\mathcal{W}}), \mathbf{u}_{\mathcal{T}b}^{\mathcal{W}})$ is the error predicted from the model $\hat{F}$ trained on the set $\mathbf{u}_{\mathcal{D}b}^{\mathcal{W}}$ and applied to set $\mathbf{u}_{\mathcal{T}b}^{\mathcal{W}}$ for classification, then the CV estimate of prediction error for a single repetition b is

$$CV_e^b(\mathbf{u}^{\mathcal{W}}) = err(\hat{F}(\mathbf{u}_{\mathcal{D}b}^{\mathcal{W}}), \mathbf{u}_{\mathcal{T}b}^{\mathcal{W}}) = \frac{1}{N_{\mathcal{T}}} \sum_{\mathbf{u}_i^{\mathcal{W}} \in \mathbf{u}_{\mathcal{T}b}^{\mathcal{W}}} Q[l_i, \eta_{\mathbf{u}_{\mathcal{D}b}^{\mathcal{W}}}(\mathbf{y}_i^{\mathcal{W}})], \tag{11}$$

and the mean CV estimate for all B repetitions is

$$MCV_e^B(\mathbf{u}^{\mathcal{W}}) = \frac{1}{B} \sum_{b=1}^{B} CV_e^b(\mathbf{u}^{\mathcal{W}}). \tag{12}$$

Let the variance of the B CV estimates be

$$VCV_e^B(\mathbf{u}^{\mathcal{W}}) = \frac{1}{B} \sum_{b=1}^{B} [CV_e^b(\mathbf{u}^{\mathcal{W}}) - MCV_e^B(\mathbf{u}^{\mathcal{W}})]^2. \tag{13}$$

From the experiments conducted, it is deduced that $VCV_e^B(\mathbf{u}^{\mathcal{Z}})$, where $\mathcal{Z} \subseteq \mathcal{W}$, depends on 1) the number of samples per emotional state N_c, 2) the number of emotional states C, and 3) $MCV_e^B(\mathbf{u}^{\mathcal{Z}})$. On the contrary, $VCV_e^B(\mathbf{u}^{\mathcal{Z}})$ does not depend on the dimensionality of the feature set $\mathcal{Z}$. In order to find a reasonable expression that correlates the three factors on which $VCV_e^B(\mathbf{u}^{\mathcal{Z}})$ depends on, three experiments are conducted.

In the first experiment, the pdfs of $f(CV_e^b(\mathbf{u}^{\mathcal{Z}}))$ for several artificially generated data sets $\mathbf{u}^{\mathcal{Z}_i}$ and $b = 1, 2, \ldots, 1000$ are estimated and plotted in Figure 2. It is inferred that $VCV_e^B(\mathbf{u}^{\mathcal{Z}})$ is inversely proportional to the number of samples per class.

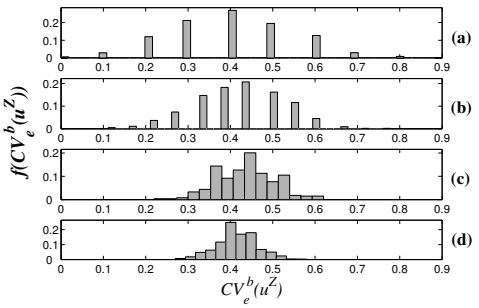

Figure 2: Pdf of $CV_e^b(\mathbf{u}^{\mathcal{Z}})$ for several feature set selections $\mathcal{Z}_i$ for N_c equal to (a) 20, (b) 36, (c) 100, and (d) 200 for 5 equiprobable classes.

In a second experiment, the modes of the pdfs of $f(CV_e^b(\mathbf{u}^{\mathcal{Z}}))$ are estimated and plotted in Figure 3 for several artificial and real data sets $\mathbf{u}^{\mathcal{Z}_i}$ and $b = 1, 2, \ldots, 1000$. The pdfs marked by $*$ correspond to three emotional speech feature sets of 5 emotional states. In each emotional state Ω_c, $N_c = 36$ utterances belong to, $c = 1, 2, \ldots, 5$. Moreover, artificially generated feature sets for five classes have been created whose prediction errors are modeled as in Figure 2. For each pdf, the peak at its mode is marked with $\circ$. It can be seen that the variance $VCV_e^B(\mathbf{u}^{\mathcal{Z}})$ depends on $MCV_e^B(\mathbf{u}^{\mathcal{Z}})$. Experimentally, it is found that $VCV_e^B(\mathbf{u}^{\mathcal{Z}})$ can be parameterized by a polynomial function of $MCV_e^B(\mathbf{u}^{\mathcal{Z}})$.

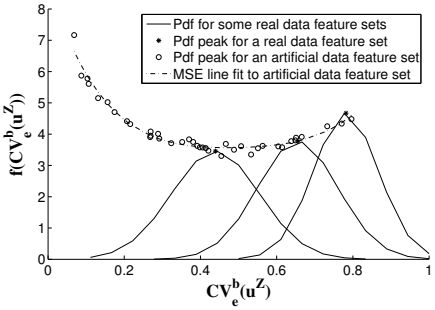

Figure 3: A parametric model for the modes of the pdf of $CV_e^b(\mathbf{u}^{\mathcal{Z}})$ for several feature sets $\mathcal{Z}_i$ selections for $C = 5$ classes with 36 samples each.

Third, by plotting the modes of $f(CV_e^b(\mathbf{u}^{\mathcal{Z}}))$ for artificially generated data sets with $N_c = 36$, $c = 2, 3, \ldots, 8$ and various $MCV_e^B(\mathbf{u}^{\mathcal{Z}})$ values in Figure 4, it is deduced that $VCV_e^B(\mathbf{u}^{\mathcal{Z}})$ is inversely proportional to C. Combining the three observations, it is found that $MCV_e^{10}(\mathbf{u}^{\mathcal{Z}})$ can be used in order to estimate $VCV_e^{\infty}(\mathbf{u}^{\mathcal{Z}})$ as follows

$$VCV_e^{\infty}(\mathbf{u}^{\mathcal{Z}}) = \frac{9.24}{\sum_{c=1}^{C} N_c} (-(MCV_e^{10}(\mathbf{u}^{\mathcal{Z}}))^2 + MCV_e^{10}(\mathbf{u}^{\mathcal{Z}})) \tag{14}$$

where the scalar value of 9.24 was found by linear regression.

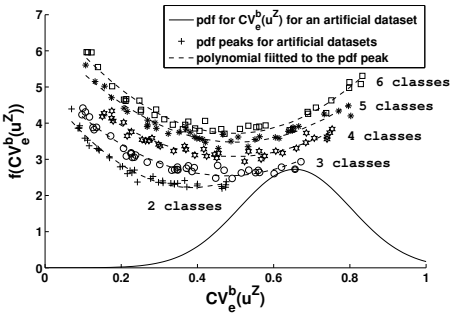

Figure 4: The modes of $f(CV_e^b(\mathbf{u}^{\mathcal{Z}}))$ for data sets that have several numbers of classes.

4. APPLICATIONS IN FEATURE SELECTION

Feature selection is used in order to determine a feature set that has the lowest classification error. We will augment the sequential floating forward selection algorithm (SFFS) by a mechanism that controls the number of crossvalidation repetitions to reduce computational burden. The SFFS consists of a forward step and a conditional backward step. The forward step is as follows. Starting from an initially empty set of features $\mathcal{Z}_0$, at each forward (inclusion) step at the level r we seek the feature $w^+ \in \mathcal{W} - \mathcal{Z}_{r-1}$ such that for $\mathcal{Z}_r = \mathcal{Z}_{r-1} \cup \{w^+\}$ the mean cross-validated error $MCV_e^{B_{thres}}(\mathbf{u}^{\mathcal{Z}_r})$ is minimized. Thus

$$w^+ = \text{argmin}_{\{w_k\} \in \mathcal{W} - \mathcal{Z}_{r-1}} [MCV_e^{B_{thres}}(\mathbf{u}^{\mathcal{Z}_{r-1} \cup \{w_k\}})] \quad (15)$$

where B_{thres} is the minimal number of crossvalidation repetitions set by the user. A typical value for B_{thres} is 50, but there is not a theoretical backround of that choice [3]. At the end of this section, an investigation on the variance of the $CV_e^b(\mathbf{u}^{\mathcal{Z}})$ will be presented, and a method to select B_{thres} will be proposed. In order to find w^+ in (15), the feature w_1 is initially registered as the feature w_{cur} which currently achieves the lowest error rate

$$J_{MinCur} = MCV_e^{B_{thres}}(\mathbf{u}^{\mathcal{Z}_{r-1} \cup \{w_1\}}) \quad (16)$$

among the non-selected features in $\mathcal{W} - \mathcal{Z}_{r-1}$. Next, w_2 is compared with w_{cur}. If $MCV_e^{B_{thres}}(\mathbf{u}^{\mathcal{Z}_{r-1} \cup \{w_2\}}) < J_{MinCur}$, then w_2 becomes w_{cur} and J_{MinCur} is set to $MCV_e^{B_{thres}}(\mathbf{u}^{\mathcal{Z}_{r-1} \cup \{w_2\}})$. Otherwise, we proceed to w_3. In general, for the kth feature $\{w_k\}$, the comparison is

$$MCV_e^{B_{thres}}(\mathbf{u}^{\mathcal{Z}_{r-1} \cup \{w_k\}}) < J_{MinCur}, \quad (17)$$

and if it is valid then $w_{cur} = w_k$.

Let us treat the error $CV_e^b(\mathbf{u}^{\mathcal{Z}})$ achieved by the Bayes classifier as a random variable. Its pdf $f(CV_e^b(\mathbf{u}^{\mathcal{Z}}))$ is a Gaussian pdf as it has been demonstrated by simulations in [8]. In inequality (17), B_{thres} CV repetitions are not necessary to see if (17) is violated. We propose to formulate a t-test in order to check whether (17) does not hold at 95% significance level for a small number of CV repetitions (e.g. B=10). If this hypothesis is accepted, the candidate feature w_k is rejected and we proceed to w_{k+1}. Otherwise, we perform B_{thres} CV repetitions and we check whether inequality (17) is valid.

In addition to the aforementioned inclusion step the SFFS algorithm applies a conditional backward step (exclusion) when no improvement can be made by any inclusion [9]. The exclusion step is as follows. We exclude at level r the $w^- \in \mathcal{Z}_r$ which achieves the highest error for the feature set $\mathcal{Z}_r - \{w^-\}$.

The $VCV_e^B(\mathbf{u}^{\mathcal{Z}})$ is of great importance when testing (17). In the forward step of feature selection algorithms, two feature sets $\mathcal{Z}_1, \mathcal{Z}_2$ must be compared in order to select the best. Let assume that $MCV_e^{B_1}(\mathbf{u}^{\mathcal{Z}_1})$ is compared against $MCV_e^{B_2}(\mathbf{u}^{\mathcal{Z}_2})$. To be certain that

$$MCV_e^{B_1}(\mathbf{u}^{\mathcal{Z}_1}) > MCV_e^{B_2}(\mathbf{u}^{\mathcal{Z}_2}), \quad (18)$$

the lower limit of the confidence interval of $MCV_e^{B_1}(\mathbf{u}^{\mathcal{Z}_1})$ should be greater than the upper limit of $MCV_e^{B_2}(\mathbf{u}^{\mathcal{Z}_2})$

$$\begin{aligned} MCV_e^{B_1}(\mathbf{u}^{\mathcal{Z}_1}) &- z_{a/2}\sqrt{VCV_e^{\infty}(\mathbf{u}^{\mathcal{Z}_1})/B_1} > \\ MCV_e^{B_2}(\mathbf{u}^{\mathcal{Z}_2}) &+ z_{a/2}\sqrt{VCV_e^{\infty}(\mathbf{u}^{\mathcal{Z}_2})/B_2}, \end{aligned} \quad (19)$$

where a=0.05 for 95% confidence intervals, and $B_1, B_2 > 30$. The unknown parameters are the number of CV repetitions B_1 and B_2. Let assume that all the confidence intervals should have the same length γ

$$\gamma = 2z_{a/2}\sqrt{VCV_e^{\infty}(\mathbf{u}^{\mathcal{Z}_i})/B_i}, \;\; i = 1, 2 \quad (20)$$

where $VCV_e^{\infty}(\mathbf{u}^{\mathcal{Z}_i})$ is estimated from the 10 CV repetitions by using (14). Then B_i can be estimated by (20) as

$$B_i = \frac{9.24(-(MCV_e^{10}(\mathbf{u}^{\mathcal{Z}_i}))^2 + MCV_e^{10}(\mathbf{u}^{\mathcal{Z}_i}))4z_{a/2}^2}{\gamma^2 \sum_{c=1}^{C} N_c}, i = 1, 2. \quad (21)$$

Subset $\mathcal{Z}_1$ is considered to be better than $\mathcal{Z}_2$ if $MCV_e^{B_1}(\mathbf{u}^{\mathcal{Z}_1}) - MCV_e^{B_1}(\mathbf{u}^{\mathcal{Z}_2}) > \gamma$. The user selects γ with respect to the computation speed, as it can be inferred from (21).

5. DECISION FUSION

The probability $p_b(y_i^{\mathcal{Z}_{opt}}|\Omega_m)$ for $\mathbf{u}_i^{\mathcal{Z}_{opt}} = (y_i^{\mathcal{Z}_{opt}}, l_i)$, where $\mathcal{Z}_{opt}$ is the optimum feature set selected by the SFFS, can be used to classify a phrase ϕ represented by the union of the utterances $\mathbf{u}_i \in \phi$. If $\phi_\rho = \bigcup_{\mathbf{u}_i \in \phi_\rho}(y_i^{\mathcal{Z}_{opt}}, l_\rho)$, where ρ is the index of the phrase, and l_ρ is the target of the lth phrase. Then the likelihood of ϕ_ρ given Ω_j is determined by

$$p(\phi_\rho|\Omega_j) = \underset{\mathbf{u}_i \in \phi_\rho}{\text{median}} \Big(\sum_{b=1}^{B_{thres}} p_b(y_i^{\mathcal{Z}_{opt}}|\Omega_j) \Big). \quad (22)$$

In (22) the median operator achieves lower error rates than the mean or the majority voting operators, because the mean is sensitive to outliers and the majority voting flattens the pdfs $p_b(y_i^{\mathcal{Z}_{opt}}|\Omega_j)$. By employing the Bayes classifier (7), then ϕ_ρ is assigned to the class with the highest probability $p(\phi_\rho|\Omega_j)$. We must note that $p(\Omega_j|\phi_\rho) = 1/C \;\; \forall j \in \{1, 2, \ldots, C\}$, because all phrases, i.e. sentences or paragraphs occur with the same frequency in DES.

6. EXPERIMENTAL RESULTS

The experiments aim at rating the discriminating capability of an optimum feature set when the proposed SFFS algorithm where the number of CV repetitions is controlled by the user is used. The data are divided according to the gender and the accent information for DES and SUSAS, respectively. In addition, to demonstrate that the utterances from paragraphs have a lower arousal level than that of words and sentences, the proposed SFFS is applied separately on Set $\mathcal{B}$ from $\mathcal{A}$. Also, a comparison of the proposed SFFS is performed against the normal SFFS for the same features and data sets. The classification errors are compared to the human error rates estimated with perception tests performed for DES in [4] and for SUSAS in [10].

As it is evident from the second and the third lines in Table 1, the proposed technique that uses the t-test to reject a feature and estimates the number of CV repetitions that should be done speeds up the execution of SFFS by 50%-60%. From the classification errors

in Table 1, we infer that there is not any significant performance deterioration between the standard algorithm and the proposed variant of SFFS. Thus the proposed SFFS is adopted throughout the remaining experiments.

A comparison of the classification error achieved by SFFS for several data sets vs. the human errors is made in Table 2. From the inspection of the second row in Table 2 we conclude that the gender information reduces classification error by 5%-7%. The classification error for set $\mathcal{B}$ is worse than that for set $\mathcal{A}$ by 7%, because the former data set is assumed to have a lower arousal, since it additionally contains utterances from long paragraphs. The classification error for the Set $\mathcal{C}$ is reduced by 2%-7% when the accent information is used.

In Table 3, the best combination of 10 features for each experiment is indicated. The energy below 250 Hz (index 86) is present in all combinations. The energy below 2100 Hz (index 90) is also quite frequent. The mean value of pitch within the rising slopes of the pitch contours (index 42), and the interquartile range of energy values (index 56) are found to be also important.

To demonstrate the usefulness of the proposed decision fusion algorithm described in section 5 we compare the classification errors measured on the sets $\mathcal{A}$ and $\mathcal{B}$ of DES with and without decision fusion. It is seen that higher errors are measured when fusion is not applied than when it does. The improvement in accuracy for the set $\mathcal{B}$ is about 7%-11%, whereas for set $\mathcal{A}$ is 1%-2%, because the number of utterances consisting a phrase in the former set is much higher than that in the latter set. The results obtained are closer to those reported for humans in the same task [4] that are listed in the last column. It is worth noting that we do not have ground truth information for emotional speech classification on utterances, whereas such ground truth is provided for emotional perception tests performed on phrases. To fill the aforementioned lack of ground truth for utterances, we assume that the latter is equal to that provided for phrases.

Experiments on set $\mathcal{B}$ are also reported in investigations [11] and [12]. The classification error is about 46% in [11], which is in agreement with our results. The only difference is that the bootstrap method was used, which is considered biased [3]. A 30% classification error is reported in [12], which is lower than the human error (33%). The low error might be due to the Fujisaki intonation parameters and the classification using only the voiced part of speech.

Table 1: Speed and performance evaluation for SFFS vs. the proposed variant.

Time lapsed (in secs)			
Method	Set $\mathcal{A}$	Set $\mathcal{B}$	Set $\mathcal{C}$
SFFS	9343	7350	8540
Proposed SFFS	4075	3270	3590
Classification error (%)			
SFFS	45.5	52.6	47.1
Proposed SFFS	46.3	53	46.3

Table 2: Classification errors on SUSAS and DES using the proposed SFFS (Mach. stands for machine and Hum. stands for Humans) .

Sets	$\mathcal{A}$	$\mathcal{A}_m$	$\mathcal{A}_f$	$\mathcal{B}$	$\mathcal{B}_m$	$\mathcal{B}_f$	$\mathcal{C}$	$\mathcal{C}_B$	$\mathcal{C}_G$	$\mathcal{C}_N$
Mac.	46.2	38	41.8	53	44.3	48.5	46.3	42.4	39	44.7
Hum.	32.7	32.4	33.1	32.7	32.4	33.1	42	41.9	40	45.4

Table 3: Best combination of features selected by the Sequential Floating Forward Selection algorithm.

Classifier	Best feature combination
Set $\mathcal{A}$	7, 10, 12, 35, 38, 56, 77, 86, 90, 111
Set $\mathcal{B}$	9, 22, 37, 39, 42, 56, 66, 76, 86, 98
Set $\mathcal{C}$	20, 30, 42, 44, 52, 65, 79, 86, 90, 109

Table 4: Classification error for the proposed SFFS without and with decision fusion.

Genders	Classification error (%)				Human errors
	without decision fusion		with decision fusion		
	Set $\mathcal{A}$	Set $\mathcal{B}$	Set $\mathcal{A}$	Set $\mathcal{B}$	
Both	46.2	53	44.5	42.8	32.7
Males	37.7	44.3	37	39.5	32.4
Females	41.8	48.5	41.5	41.2	33.1

7. CONCLUSIONS

First, we have described how sequential floating forward feature selection algorithm can be accelerated. The proposed method can be applied to other subset selection algorithms such as the branch and bound or the backward selection. The second contribution of the paper was in the combination of partial emotional speech classification decisions from short speech segments in order to derive a unique, more robust, decision on the basis of long phrases. When gender and accent information is taken into account the reported errors are approaching the human errors.

REFERENCES

[1] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Communication*, vol. 40, pp. 227–256, 2003.

[2] T. Vogt and E. André, "Comparing feature sets for acted and spontaneous speech inview of automatic emotion recognition," in *Proc. Int. Conf. Multimedia & Expo*, Amsterdam, 2005.

[3] B. Efron and R. E. Tibshirani, *An Introduction to the Bootstrap*, N.Y.: Chapman & HALL/CRC, 1993.

[4] I. S. Engberg and A. V. Hansen, "Documentation of the Danish Emotional Speech database (DES)," Internal report, Center for Person Kommunikation, Aalborg University, 1996.

[5] J. H. L. Hansen, "Analysis and compensation of speech under stress and noise for enviromental robustness in speech recognition," *Speech Communication*, vol. 20, pp. 151–173, 1996.

[6] R. Cowie and E. Douglas-Cowie, "Automatic statistical analysis of the signal and prosodic signs of emotion in speech," in *Proc. Int. Conf. Spoken Language Processing*, 1996, vol. 3, pp. 1989–1992.

[7] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discete-Time Processing of Speech Signals*, N.Y.: Wiley & Sons, 2000.

[8] D. Ververidis and C. Kotropoulos, "Sequential forward feature selection with low computational cost," in *Proc.* XIII *European Signal Processing Conf.*, Antalya, Turkey, 2005.

[9] P. Pudil, J. Novovicova, and J. Kittler, "Floating search methods in feature selection," *Pattern Recognition Letters*, vol. 15, pp. 1119–1125, 1994.

[10] R. S. Bolia and R. E. Slyh, "Perception of stress and speaking style for selected elements of the SUSAS database," *Speech Communication*, vol. 40, pp. 493–501, 2003.

[11] Z. Hammal, B. Bozkurt, L. Couvreur, D. Unay, A. Caplier, and T. Dutoit, "Passive versus active: vocal classification system," in *Proc.* XIII *European Signal Processing Conference*, Antalya, Turkey, 2005.

[12] P. Zervas, I. Mporas, N. Fakotakis, and G. Kokkinakis, "Employing Fujisaki 's intonation model parameters for emotion recognition," in *Proc. 4th Hellinic Conf. Artificial Intelligence (SETN'06)*, Heraklion, Crete, May 2006.