

# Gender Classification in Emotional Speech

Mohammad Hossein Sedaaghi  
*Sahand University of Technology*  
*Iran*

## 1. Introduction

The emotion accompanying with the voice is considered as a salient aspect in human communication. The effects of emotion in speech tend to alter the voice quality, timing, pitch and articulation of the speech signal. Gender classification, on the other hand, is an interesting field for psychologists to foster human-technology relationships. Automatic gender classification take on an increasingly ubiquitous role in myriad of applications, e.g., demographic data collection. An automatic gender classifier assists the development of improved male and female voice synthesizers (Childers et. al., 1988). Gender classification is also used to improve the speaker clustering task which is useful in speaker recognition. By separately clustering each gender class, the search space is reduced when evaluating the proposed hierarchical agglomerative clustering algorithm (Tranter and Reynolds, 2006). It also avoids segments having opposite gender tags being erroneously clustered together. Gender information is time-invariant, phoneme-independent, and identity-independent for speakers of the same gender (Wu & Childers, 1991). In (Xiaofan & Simske, 2004), an accent classification method is introduced on the top of gender classification. Vergin et al. (Vergin, 1996) claim that the use of gender-dependent acoustic-phonetic models reduces the word error rate of the baseline speech recognition system by 1.6%. In (Harb & Chen, 2005), a set of acoustic and pitch features along with different classifiers is tested for gender identification. The fusion of features and classifiers is shown to perform better than any individual classifier. A gender classification system is proposed in (Zeng et. al., 2006) based on Gaussian mixture models of speech features. Metze et al. have compared four approaches for age and gender recognition using telephone speech (Metze et. al., 2007). Gender cues elicited from the speech signal are useful in content-based multimedia indexing as well (Harb & Chen, 2005). Gender-dependent speech emotion recognizers have been shown to perform better than gender-independent ones for five emotional state (Ververidis & Kotropoulos, 2004; Lin & Wei, 2005) in DES (Engberg & Hansen, 1996). However, gender information is taken for granted there. The most closely related work to the present one is related to the research by Xiao et al. (Xiao et. al., 2007), where gender classification was incorporated in emotional speech recognition system using a wrapper approach based on back-propagation neural networks with sequential forward selection. An accuracy of 94.65% was reported for gender classification on the Berlin dataset (Burkhardt et. al., 2005). In this research, we employ several classifiers and assess their performance in gender classification by processing utterances from DES (Engberg & Hansen, 1996), SES (Sedaaghi, 2008) and GES (Burkhardt et. al., 2005) databases. They all contain affective speech. In

particular, we test the Bayes classifier with sequential floating forward feature selection (SFFS) (Fukunaga & Narendra, 1975; Pudil et. al., 1994), the probabilistic neural networks (Specht, 1990), the support vector machines (Vapnik, 1998), and the  $K$ -nearest neighbor classifiers (Fix & Hodges, 1991-a; Fix & Hodges, 1991-b). Although techniques based on hidden Markov models could be applied for gender classification in principle, they are not included in this study, because temporal information is ignored.

## 2. Database

The first dataset stems from Danish Emotional Speech (DES) database, which is publicly available and well annotated (Engberg & Hansen, 1996). The recordings in DES include utterances expressed by two professional actors and two actresses in five different emotional states (anger, happiness, neutral, sadness, and surprise). The utterances correspond to isolated words, sentences, and paragraphs. The complete database comprise approximately 30 minutes of speech.

Sahand Emotional Speech (SES) database (Sedaaghi, 2008) comprise utterances expressed by five male and five female students in five emotional states similar to the emotions employed in DES. Twenty four words, short sentences and paragraphs spoken in Farsi by each student are included in SES database leading to 1200 utterances and about 50 minutes recording.

As the third database, the database of German Emotional Speech (GES) is investigated. An emotional database comprising 6 basic emotions (anger, joy, sadness, fear, disgust and boredom) as well as neutral speech is recorded (Burkhardt et. al., 2005). Ten professional native German actors (5 female and 5 male) have simulated these emotions, producing 10 utterances (5 short and 5 longer sentences). The recorded speech material of about 800 sentences have been evaluated with respect to recognizability and naturalness in a forced-choice automated listening-test by 20-30 judges. Those utterances for which the emotion is recognized by at least 80% of the listeners are used for further analysis (i.e., 535 sentences) (Burkhardt et. al., 2005).

## 3. Feature extraction

The automatic gender classification is mainly achieved based on the average value of the fundamental frequency (i.e.,  $F_0$ ). Also, the distinction between men and women have been represented by the location in the frequency domain of the first 3 formants for vowels (Peterson & Barney, 1952). To improve the efficiency, more features should be considered. The statistical features employed in our study are grouped in several classes and have been demonstrated in Table 1. They have been adopted from (Ververidis & Kotropoulos, 2006).

	<b>Formant features</b>
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.
	<b>Pitch features</b>
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.
	<b>Intensity (Energy) features</b>
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.
	<b>Spectral features</b>
86-93	Energy below 250, 600, 1000, 1500, 2100, 2800, 3500, 3950 Hz.
94-100	Energy in the frequency bands 250-600, 600-1000, 1000-1500, 1500-2100, 2100-2800, 2800-3500, 3500-3950 Hz.
101-106	Energy in the frequency bands 250-1000, 600-1500, 1000-2100, 1500-2800, 2100-3500, 2800-3950 Hz.
107-111	Energy in the frequency bands 250-1500, 600-2100, 1000-2800, 1500-3500, 2100-3950 Hz.
112-113	Energy ratio between the frequency bands (3950-2100) and (2100-0) and between the frequency bands (2100-1000) and (1000-0).

Table 1. List of extracted features adopted from (Ververidis &amp; Kotropoulos, 2006).

Not all the features can be extracted from each utterance. For example, some pitch contours do not have plateaux below 45% of their maximum pitch value, or some utterances do not have pitch at all because they are unvoiced. When a large number of missing feature values is met, the corresponding feature is discarded. The features with NaN (not a number) values are replaced with the mean value of the corresponding feature. The outliers (features with value 10000 times greater or smaller than the median value) are then eliminated. Also the features with bias are investigated. Then all features are normalized. The discarded features are as follows.

- DES: 8, 17-51, 57-85, 105 (47 features remained),
- SES: 8, 23-29, 33-34, 41, 48, 57-63, 67, 75, 82, 94, 96, 98, 103-105, 109-113 (80 features preserved),
- GES: 8, 23-29, 33-34, 41, 60, 67, 75, 82, 94, 96, 98-99, 103-107, 109-113 (84 features retained).

#### 4. Classifiers

The output of the gender classifier on emotional speech is a prediction value (label) of the actual speaker's gender. In order to evaluate the performance of a classifier, the repeated s-fold cross-validation method is used. According to this method if  $s=20$ , the utterances in the data collection are divided into a training set containing 80% of the available data and a disjoint test set containing the remaining 20% of the data. The procedure is repeated for  $s=20$  times. The training and the test set are selected randomly. The classifier is trained using the training set and the classification error is estimated on the test set. The estimated classification error is the average classification error over all repetitions (Efron & Tibshirani, 1993).

The following classifiers have been investigated:

1. Naive Bayes classifier using the SFFS feature selection method (Pudil et. al., 1994). The SFFS consists of a forward (inclusion) step and a conditional backward (exclusion) step that partially avoids local optima. In the proposed method, feature selection is used in order to determine a set of 20 features that yields the lowest prediction error for a fixed number of cross-validation repetitions. Ten best sorted features among the 20 best selected features are as follows.
  - 10 best features for DES: {112, 15, 10, 107, 96, 52, 102, 14, 13, 99},
  - 10 best features for SES: {6, 32, 51, 3, 76, 20, 44, 52, 17, 22},
  - 10 best features for GES: {38, 69, 43, 80, 42, 40, 63, 8, 15, 6}.
2. Probabilistic Neural Networks (PNNs) (Specht, 1990). PNNs are a kind of radial basis function (RBF) networks suitable for classification problems. A PNN employs an input, a hidden, and an output layer. The input nodes forward the values admitted by patterns to the hidden layer ones. The hidden layer nodes are as many as the input nodes. They are simply RBFs that nonlinearly transform pattern values to activations. The nodes at the output layer are as many as the classes. Each node sums the activation values weighted possibly by proper weights. The input pattern is finally classified to the class associated to the output node whose value is maximum. PNNs with a spread parameter equal to 0.1 are found to yield the best results. If the spread parameter is near zero, the network acts as a nearest neighbor classifier. As the spread parameter becomes large, the network takes into account several nearby patterns.

3. Support vector machines (SVMs) (Vapnik, 1998). SVMs with five different kernels, have been used. Training was performed by the least-squares method. The following kernel functions have been tested:
  - Gaussian RBF (denoted SVM1):  $K(x_i, x_j) = \exp\{-\gamma \|x_i - x_j\|^2\}$  with  $\gamma = 1$ ;
  - multilayer perceptron (denoted SVM2):  $K(x_i, x_j) = S(x_i^T x_j - 1)$ , where  $S(\cdot)$  is a sigmoid function;
  - Quadratic kernel (denoted SVM3):  $K(x_i, x_j) = (x_i^T x_j + 1)^2$ ;
  - Linear kernel (denoted SVM4):  $K(x_i, x_j) = x_i^T x_j$ ;
  - Polynomial kernel (denoted SVM5):  $K(x_i, x_j) = (x_i^T x_j + 1)^3$ . A polynomial kernel of degree 4 is found to yield the same results with the cubic kernel.
4. For  $K$ -NNs, it is hard to find systematic methods for selecting the optimum number of the closest neighbors and the most suitable distance. Four  $K$ -NNs have been employed with different distance functions, such as the Euclidean distance denoted as KNN1, cityblock (i.e., sum of absolute differences) denoted as KNN2, cosine-based (i.e. one minus the cosine of the included angle between patterns) denoted as KNN3 and correlation-based (i.e., one minus the sample correlation between patterns) denoted as KNN4, respectively. We have selected  $K=2$  in all experiments. Other values of  $K$  did not affect the classification accuracy unless the consensus rule was applied instead of the normal rule. In this case, none of the results of the  $K$ -NN would be stable and thus valid for classification.
5. Gaussian Mixture model (GMM) have been employed in many fields, e.g., speech and speaker recognition (Stephen & Paliwal, 2006; Reynolds & Rose, 1995). In GMM, during the training phase, pdf (probability density function) parameters for each class (gender) are estimated. Then, during the classification phase, a decision is taken for each test utterance by computing the maximum likelihood criterion. GMM is a combination of  $K$  Gaussian laws. Each law in the mixture is weighted and specified by two parameters: the mean and the covariance matrix ( $\Sigma_k$ ).

## 5. Comparative results

Figure 1 illustrates the correct classification rates achieved by each of the aforementioned 11 classifiers on DES database, when 20% of the total utterances have been used for testing. For each classifier, columns "Total", "Male", and "Female" correspond to the total correct classification rate, the rate of correct matches between the actual gender and the predicted one by the classifier for utterances uttered by male speakers, and the rate of correct matches between the actual gender and the predicted one by the classifier for utterances uttered by female speakers, respectively. The leftmost column shows the total correct classification rate. The middle and the rightmost columns are the classification rates that correspond to correct matches between the actual speaker gender (i.e. the ground truth) and the gender prediction by the classifier for male and female speakers, separately. In the sequel, the total correct classification rate, the correct classification rate for male speakers, and the correct classification rate for female speakers are abbreviated as TCCR, MCCR, and FCCR, respectively. In Figure 1, the maximum and minimum TCCR for DES were obtained by the SVM1 (90.94%) and the SVM2 (57.33%), respectively. The maximum and minimum MCCR for DES were related to GMM (95.42%) and SVM2 (58.11%), respectively. For FCCR on DES, the maximum and minimum values were obtained by the Bayes classifier with SFFS (91.07%) and SVM2 (56.54%), respectively. The best results for TCCR, MCCR and FCCR are marked with "↓" sign.

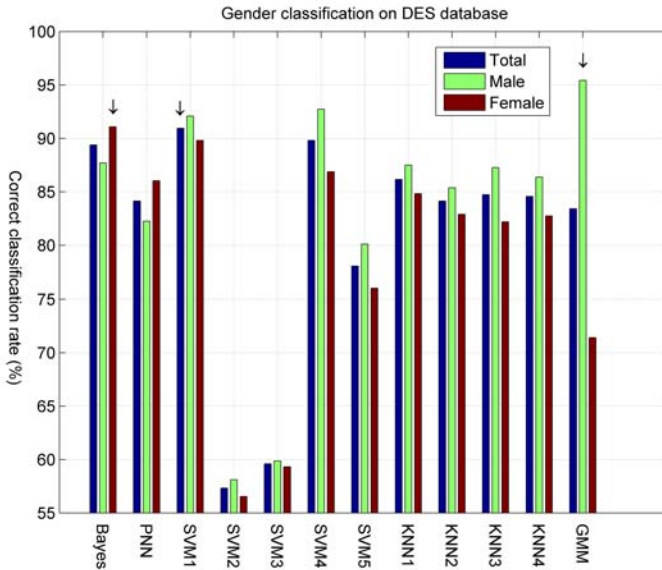


Fig. 1. Correct classification rates on DES database for the different methods when the size of test utterances is 20% of the total utterances.

Figures 2 & 3 demonstrate similar results for SES and GES databases, respectively.

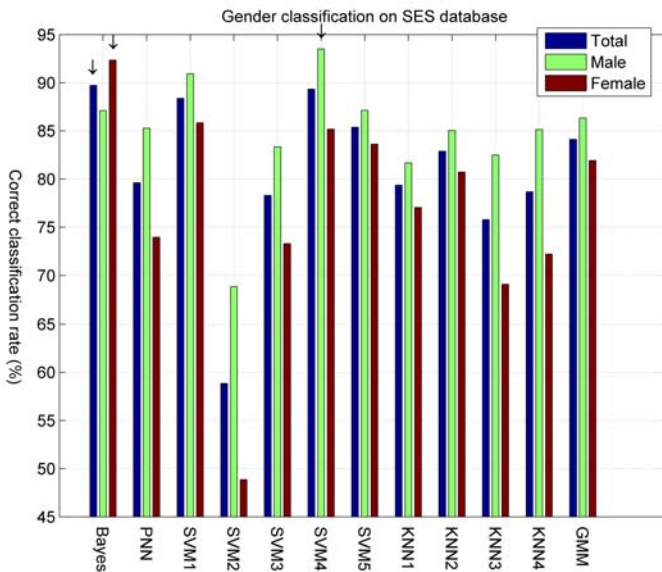


Fig. 2. Correct classification rates on SES database for the different methods when the size of test utterances is 20% of the total utterances.

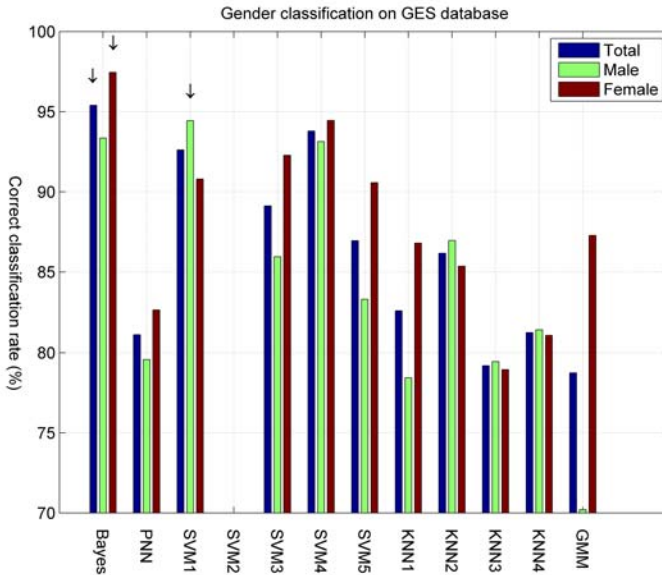


Fig. 3. Correct classification rates on GES database for the different methods when the size of test utterances is 20% of the total utterances.

In Figure 2, the maximum and minimum TCCR for SES were obtained by the the Bayes classifier using SFFS (89.73%) and the SVM2 (58.83%), respectively. The maximum and minimum MCCR for SES were related to SVM4 (93.51%) and SVM2 (68.83%), respectively. For FCCR on SES, the maximum and minimum values were obtained by the Bayes classifier with SFFS (92.36%) and SVM2 (48.86%), respectively.

In Figure 3, the maximum and minimum TCCR for GES were obtained by Bayes+SFFS (95.40%) and the GMM (78.74%), respectively. This is where SVM2 failed to classify at all. The maximum and minimum MCCR for GES were related to SVM1 (94.43%) and GMM (70.20%), respectively. The maximum and minimum values for FCCR on GES, were achieved by the Bayes classifier with SFFS (97.45%) and KNN3 (78.94%), respectively.

In the following, we concentrate on the top methods, i.e., SVM1, SVM4, GMM, and the Bayes classifier with SFFS. Table 2 demonstrates the confusion matrix for gender classification of the top methods after running each method several times and taking the mean value. The correct classification rates for each gender are shown in boldface. SVM1 outperforms the other methods achieving a correct classification rate of 90.94% (TCCR) with a standard deviation of 0.65. GMM is the best classifier, when the correct matches are between the actual gender and the predicted one by the classifier are measured for actors' utterances, yielding a rate of 95.42% (MCCR). The Bayes classifier using SFFS achieves a rate of 91.07%, when the correct matches between the actual gender and the predicted one by the classifier are measured for actresses' utterances (FCCR).

Similarly, Tables 3 & 4 show the confusion matrices for gender classification of the top methods on SES and GES databases, respectively. The Bayes classifier using SFFS outperforms the other methods achieving a correct classification rate of 89.74% (TCCR) with a standard deviation of 0.103 on SES. It is also the best classifier for FCCR with 92.36% on

SES. SVM4 is considered as the best classifier for MCCR with 93.51% on SES. Also Bayes classifier using SFFS outperforms other classifiers for TCCR with 95.40% on GES with a standard deviation of 1.16. Moreover, it is the best classifier for FCCR with 97.45% on GES. SVM1 is the best classifier for MCCR with 94.43% on GES.

<b>GMM</b>		Response (%)		<b>Bayes-SFFS</b>		Response (%)	
Ground Truth ↓		Male	Female	Ground Truth ↓		Male	Female
Male		<b>95.42</b>	4.58	Male		<b>87.69</b>	12.31
Female		28.59	<b>71.41</b>	Female		8.93	<b>91.07</b>
Correct rate		83.42%		Correct rate		89.38%	
<b>SVM1</b>		Response (%)		<b>SVM4</b>		Response (%)	
Ground Truth ↓		Male	Female	Ground Truth ↓		Male	Female
Male		<b>92.08</b>	7.92	Male		<b>92.72</b>	7.28
Female		10.19	<b>89.81</b>	Female		13.12	<b>86.88</b>
Correct rate		90.95%		Correct rate		89.80 %	

Table 2. Confusion matrix for the 4 best methods when 20% of the utterances of DES database are used for testing.

<b>GMM</b>		Response (%)		<b>Bayes-SFFS</b>		Response (%)	
Ground Truth ↓		Male	Female	Ground Truth ↓		Male	Female
Male		<b>86.34</b>	13.66	Male		<b>87.11</b>	12.89
Female		18.08	<b>81.92</b>	Female		7.64	<b>92.36</b>
Correct rate		84.13%		Correct rate		89.74%	
<b>SVM1</b>		Response (%)		<b>SVM4</b>		Response (%)	
Ground Truth ↓		Male	Female	Ground Truth ↓		Male	Female
Male		<b>90.94</b>	9.06	Male		<b>93.51</b>	6.49
Female		14.16	<b>85.84</b>	Female		14.83	<b>85.17</b>
Correct rate		88.39%		Correct rate		89.34%	

Table 3. Confusion matrix for the 4 best methods when 20% of the utterances of SES database are used for testing.

<b>GMM</b>		Response (%)		<b>Bayes-SFFS</b>		Response (%)	
Ground Truth ↓		Male	Female	Ground Truth ↓		Male	Female
Male		<b>70.20</b>	29.80	Male		<b>93.34</b>	6.66
Female		12.72	<b>87.28</b>	Female		2.55	<b>97.45</b>
Correct rate		78.74%		Correct rate		95.40%	
<b>SVM1</b>		Response (%)		<b>SVM4</b>		Response (%)	
Ground Truth ↓		Male	Female	Ground Truth ↓		Male	Female
Male		<b>94.43</b>	5.57	Male		<b>93.13</b>	6.87
Female		9.21	<b>90.79</b>	Female		5.55	<b>94.45</b>
Correct rate		92.61%		Correct rate		93.79%	

Table 4. Confusion matrix for the 4 best methods when 20% of the utterances of GES database are used for testing.

In the following, the behaviour of the best classifiers are investigated against changing the parameters. Figures 4, 5 & 6 highlight the behaviour of the Bayes classifier with SFFS on DES, SES and GES databases, respectively, for varying numbers of cross-validation repetitions and varying portions of utterances engaged in testing. The flatness of the shapes confirms that if we select 20% of the utterances for testing and 20 repetitions, our judgements are fair.

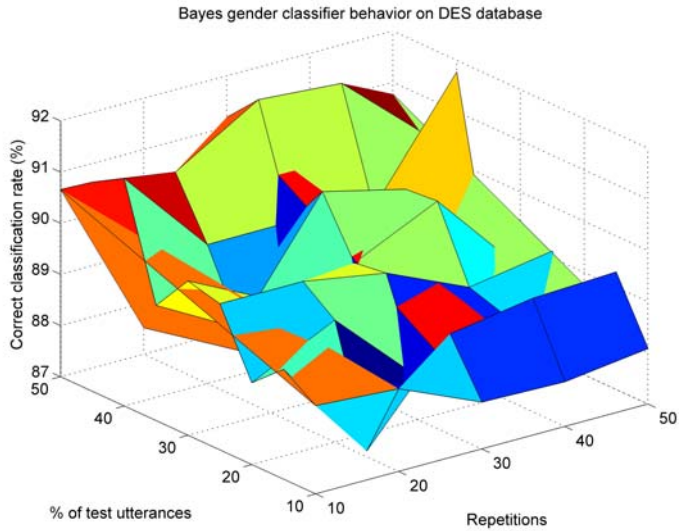


Fig. 4. Probability of correct classification of the Bayes classifier with SFFS on DES database for varying repetitions and portions of the utterances used during testing.

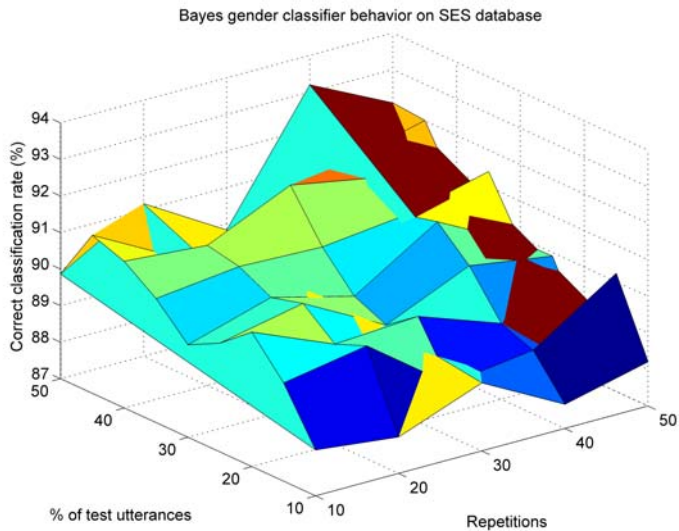


Fig. 5. Probability of correct classification of the Bayes classifier with SFFS on SES database for varying repetitions and portions of the utterances used during testing.

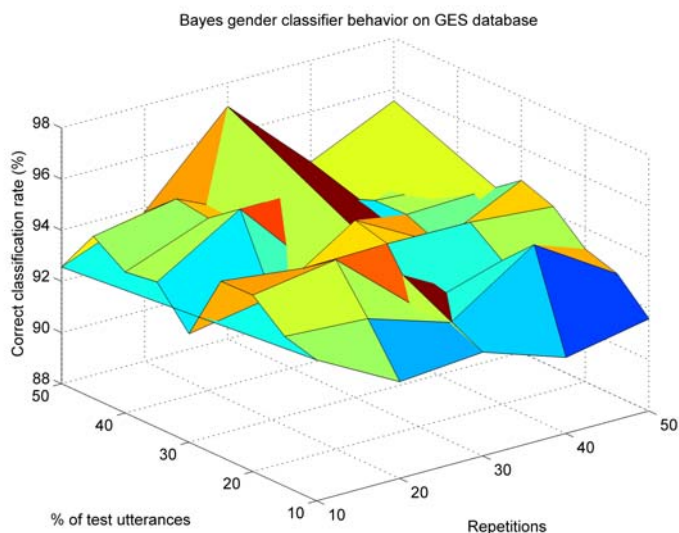


Fig. 6. Probability of correct classification of the Bayes classifier with SFFS on GES database for varying repetitions and portions of the utterances used during testing.

Tables 5, 6 and 7 investigate, in detail, the minimum and maximum rates measured for the Bayes classifier with SFFS on DES, SES and GES databases, respectively. The minimum TCCR for DES, SES and GES was measured when 20, 40 and 40 repetitions were made using 15%, 10% and 45% of utterances for testing, respectively. The maximum TCCR for DES, SES and GES was measured by making 30, 40 and 30 repetitions and employing 45%, 50% and 50% of the available utterances for testing, respectively. The minimum MCCR for DES, SES and GES was measured when 50, 40 and 40 repetitions were made while using 30%, 10% and 45% of utterances for testing, respectively. The maximum MCCR for DES, SES and GES was measured by making 40, 50 and 30 repetitions and employing 45%, 50% and 50% of the available utterances for testing, respectively. For FCCR on DES, SES and GES, 20, 10 and 40 repetitions and 50%, 50% and 45% of utterances for testing yield the minimum rate, respectively, while 30, 40 and 20 repetitions and 45%, 50% and 50% of the utterances engaged in testing are required for the maximum rate, respectively.

Rates	Min (%)	Max (%)	Mean (%)	Std (%)
TCCR	87.11	91.82	89.46	1.22
MCCR	83.94	92.58	87.71	1.71
FCCR	87.07	93.75	91.21	1.49

Table 5. Behaviour of Bayes classifier with SFFS for gender classification on DES database.

Rates	Min (%)	Max (%)	Mean (%)	Std (%)
TCCR	87.66	93.19	90.10	1.22
MCCR	83.96	90.28	87.42	1.55
FCCR	89.83	96.49	92.79	1.52

Table 6. Behaviour of Bayes classifier with SFFS for gender classification on SES database.

Rates	Min (%)	Max (%)	Mean (%)	Std (%)
TCCR	89.67	97.08	93.69	1.34
MCCR	85.66	97.74	90.59	2.25
FCCR	93.67	99.47	96.79	1.21

Table 7. Behaviour of Bayes classifier with SFFS for gender classification on GES database.

Tables 8-10 illustrate the behaviour of SVM1 on DES, SES and GES databases, respectively, when the size of the test utterances ranges between 10% and 50% of the available utterances. For TCCR on DES, SES and GES, 50%, 45% and 45% of the available utterances yield the minimum value, while 40%, 10% and 10% of the utterances yield the maximum value, respectively. For MCCR, 25%, 50% and 35% the test utterances yield the minimum value while 40%, 10% and 25% of the available utterances yield the maximum value for MCCR. For FCCR, 15%, 45% and 45% of the utterances engaged during testing yield the minimum value, while 20%, 15% and 15% of the utterances yield the maximum value.

Tables 11 and 12 show the behaviour of SVM4 on DES, SES and GES databases. The size of the test utterances ranges between 10% and 50% of the available utterances. For TCCR on DES, SES and GES, 30%, 35% and 50% of the available utterances yield the minimum value,

Rates	Min (%)	Max (%)	Mean (%)	Std (%)
TCCR	88.91	91.17	89.79	0.82
MCCR	89.45	94.14	91.38	1.50
FCCR	86.18	89.81	88.19	1.32

Table 8. Behaviour of SVM1 on DES database for gender classification when the size of the test utterances varies between 10% and 50% of the utterances.

Rates	Min (%)	Max (%)	Mean (%)	Std (%)
TCCR	86.14	89.71	87.59	1.25
MCCR	88.32	93.14	90.21	1.45
FCCR	82.97	86.37	84.96	1.31

Table 9. Behaviour of SVM1 on SES database for gender classification when the size of the test utterances varies between 10% and 50% of the utterances.

Rates	Min (%)	Max (%)	Mean (%)	Std (%)
TCCR	89.49	92.93	91.32	1.23
MCCR	90.73	95.41	93.01	2.01
FCCR	87.79	91.48	89.64	1.23

Table 10. Behaviour of SVM1 on GES database for gender classification when the size of the test utterances varies between 10% and 50% of the utterances.

Rates	Min (%)	Max (%)	Mean (%)	Std (%)
TCCR	87.19	90.46	88.68	1.19
MCCR	88.19	92.72	90.40	1.39
FCCR	83.81	89.17	86.96	1.78

Table 11. Behaviour of SVM4 on DES database for gender classification when the size of the test utterances varies between 10% and 50% of the utterances.

Rates	Min (%)	Max (%)	Mean (%)	Std (%)
TCCR	86.87	89.56	88.09	0.93
MCCR	91.05	94.09	92.75	0.98
FCCR	80.17	86.23	83.44	2.03

Table 12. Behaviour of SVM4 on SES database for gender classification when the size of the test utterances varies between 10% and 50% of the utterances.

Rates	Min (%)	Max (%)	Mean (%)	Std (%)
TCCR	90.30	93.91	92.38	1.32
MCCR	87.86	93.88	91.07	2.10
FCCR	92.72	94.69	93.68	0.76

Table 13. Behaviour of SVM4 on GES database for gender classification when the size of the test utterances varies between 10% and 50% of the utterances.

while 15%, 25% and 10% of the utterances yield the maximum value, respectively. For MCCR, 30%, 15% and 50% the test utterances yield the minimum value while 20%, 10% and 15% of the available utterances yield the maximum value for MCCR. For FCCR, 10%, 35% and 25% of the utterances engaged during testing yield the minimum value, while 40%, 15% and 10% of the utterances yield the maximum value.

Tables 14-16 illustrate the behaviour of GMM on DES and SES databases, respectively, when the size of the test utterances ranges between 10% and 50% of the available utterances (GMM is not a good classifier for GES). For TCCR on DES and SES, 40% and 50% of the available utterances yield the minimum value, while 50% and 10% of the utterances yield the maximum value, respectively. For MCCR, 50% and 20% of the test utterances yield the minimum value while 10% and 50% of the available utterances yield the maximum value for MCCR. For FCCR, 40% and 50% of the utterances engaged during testing yield the minimum value, while 50% and 10% of the utterances yield the maximum value.

Rates	Min (%)	Max (%)	Mean (%)	Std (%)
TCCR	82.02	84.81	83.59	0.89
MCCR	91.73	96.06	94.71	1.33
FCCR	68.88	77.89	72.47	2.70

Table 14. Behaviour of GMM on DES database for gender classification when the size of the test utterances varies between 10% and 50% of the utterances.

Rates	Min (%)	Max (%)	Mean (%)	Std (%)
TCCR	69.66	85.03	79.97	5.53
MCCR	86.34	92.49	88.92	2.37
FCCR	46.83	83.67	71.02	13.36

Table 15. Behaviour of GMM on SES database for gender classification when the size of the test utterances varies between 10% and 50% of the utterances.

The computational speed was measured using a PC P4, 3GHz CPU and 1 GB RAM while a virus shield was active.

Classifier	DES	SES	GES
GMM	21.81	31.80	44.36
Bayes+SFFS	30.22	48.11	30.76
PNN	7.88	1.33	1.98
SVM1	2.90	1.78	0.34
SVM2	1.62	1.73	0.33
SVM3	1.34	1.46	0.28
SVM4	1.30	1.43	0.28
SVM5	1.38	1.47	0.28
KNN1	0.87	1.00	0.30
KNN2	0.69	0.99	0.30
KNN3	0.40	0.41	0.16
KNN4	0.44	0.42	0.18

Table 17. Computational time (in sec) for different classifiers on different databases.

Accordingly, SVM1 outperforms the other methods with respect to all the four factors: TCCR, MCCR, FCCR, and speed for emotional speech.

However, for non-emotional speech, we recommend GMM.

## 6. Conclusions

We have investigated several popular methods for gender classification by processing emotionally colored speech from the DES, SES and GES databases. Based on the results, several conclusions can be drawn. The SVM with a Gaussian RBF kernel (SVM1) has demonstrated to yield the most accurate results considering other parameters such as the computation speed. The correct gender classification rates have been more than 90% when emotional speech utterances from both genders were processed, or when emotional speech utterances of male or female speakers were used. Another acceptable alternative is the Bayes classifier using sequential floating forward feature selection.

## 7. References

- F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier and B. Weiss (2005). A database of German Emotional Speech. In Proc. Interspeech 2005 Conf. Lisbon, Portugal.
- D. G. Childers, K. Wu, and D. M. Hicks (1987). Factors in voice quality: acoustic features related to gender. In Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, volume 1, pages 293–296.
- B. Efron and R. E. Tibshirani (1993), *An Introduction to the Bootstrap*, Chapman & Hall/CRC, N.Y..
- I. S. Engberg and A. V. Hansen (1996). Documentation of the Danish Emotional Speech database (DES). Technical Report Internal AAU report, Center for Person, Kommunikation, Aalborg Univ., Denmark.
- E. Fix and J. Hodges (1991-a). Discriminatory analysis, nonparametric discrimination, consistency properties. In B. Dasarthy, editor, *Nearest Neighbor Pattern Classification Techniques*, pages 32–39. IEEE Computer Society Press, Los Alamitos, CA.
- E. Fix and J. Hodges (1991-b). Discriminatory analysis: small sample performance. In B. Dasarthy, editor, *Nearest Neighbor Pattern Classification Techniques*, pages 40–56. IEEE Computer Society Press, Los Alamitos, CA.

- K. Fukunaga and P. M. Narendra (1975). A branch and bound algorithm for computing k-nearest neighbors. *IEEE Trans. Computers*, 24:750-753.
- H. Harb and L. Chen (2005). Voice-based gender identification in multimedia applications. *J. Intelligent Information Systems*, 24(2):179-198.
- Y. L. Lin and G. Wei (2005). Speech emotion recognition based on HMM and SVM. In *Proc. IEEE Int. Conf. Machine Learning and Cybernetics*, volume 8, pages 4898-4901. Guangzhou, China.
- F. Metze, J. Ajmera, R. Englert, U. Bub, F. Burkhardt, J. Stegmann, C. Muller, R. Huber, B. Andrassy, J. G. Bauer, and B. Little (2007). Comparison of four approaches to age and gender recognition for telephone applications. In *Proc. 2007 IEEE Int. Conf. Acoustics, Speech and Signal Processing*, volume 4, pages 1089-1092. Honolulu.
- G. Peterson and H. Barney (1952). Control methods used in a study of vowels. *Journal of Acoustical Society of America*, 24, 175-184.
- P. Pudil, J. Novovicova, and J. Kittler (1994). Floating search methods in feature selection. *Pattern Recognition Letters*, 15(11):1119-1125.
- D. Reynolds and R. Rose (1995). Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, vol. 3(1): 72-83.
- M. H. Sedaaghi (2008). Documentation of the Sahand Emotional Speech database (SES). Technical Report, Department of Electrical Eng., Sahand Univ. of Tech, Iran.
- D. F. Specht (1990). Probabilistic neural networks. *Neural Networks*, 3:109-118.
- S. Stephen and K. K. Paliwal (2006). Scalable distributed speech recognition using Gaussian mixture model-based block quantisation, *Speech Communication*, vol. 48: 746-758.
- S.E. Tranter and D. A. Reynolds (2006). An Overview of Automatic Speaker Diarisation Systems. *IEEE Trans. Speech & Audio Processing: Special issue on Rich Transcription*, 14(5): 1557-1565.
- V. N. Vapnik (1998). *The Nature of Statistical Learning Theory*. Springer, N.Y..
- R. Vergin, A. Farhat, and D. O'Shaughnessy (1996). Robust gender-dependent acoustic phonetic modelling in continuous speech recognition based on a new automatic male/female classification. In *Proc. Int. IEEE Conf. Acoustics, Speech, and Signal Processing (ICASSP-96)*, volume 2, pages 1081-1084. Atlanta.
- D. Ververidis and C. Kotropoulos (2004). Automatic speech classification to five emotional states based on gender information. In *Proc. European Signal Processing Conf. (EUSIPCO 04)*, volume 1, pages 341-344. Vienna, Austria.
- D. Ververidis and C. Kotropoulos (2006). Fast sequential floating forward selection applied to emotional speech features estimated on DES and SUSAS data collections, in *Proc. 14th. European Signal Processing Conf. Florence, Italy*.
- K. Wu and D. G. Childers (1991). Gender recognition from speech. Part I: Coarse analysis. *J. Acoust. Soc. of Am.*, 90(4):1828-1840.
- Z. Xiao, E. Dellandrea, W. Dou, and L. Chen (2007). Hierarchical classification of emotional speech. Technical Report RR-LIRIS-2007-06, LIRIS UMR 5205 CNRS.
- L. Xiaofan and S. Simske (2004). Phoneme-less hierarchical accent classification. In *Proc. 38th. Asilomar Conf. Signals, Systems and Computers*, volume 2, pages 1801-1804. California.
- Y. Zeng, Z. Wu, T. Falk, and W. Y. Chan (2006). Robust GMM based gender classification using pitch and RASTA-PLP parameters of speech. In *Proc. 5th. IEEE Int. Conf. Machine Learning and Cybernetics*, pages 3376-3379. China.