Gender Classification In Two Emotional Speech Databases

Margarita Kotti and Constantine Kotropoulos
Department of Informatics, Aristotle Univ. of Thessaloniki Box 451, Thessaloniki 541 24, Greece
{*mkotti,costas*} *@aiia.csd.auth.gr*

Abstract

Gender classification is a challenging problem, which finds applications in speaker indexing, speaker recognition, speaker diarization, annotation and retrieval of multimedia databases, voice synthesis, smart human-computer interaction, biometrics, social robots etc. Although it has been studied for more than thirty years, by no means it is a solved problem. Processing emotional speech in order to identify speaker's gender makes the problem even more interesting. A large pool of 1379 features is created including 605 novel features. A branch and bound feature selection algorithm is applied to select a subset of 15 features among the 1379 originally extracted. Support vector machines with various kernels are tested as gender classifiers, when applied to two databases, namely: the Berlin database of Emotional Speech and the Danish Emotional Speech database. The reported classification results outperform those obtained by state-of-the-art techniques, since a perfect classification accuracy is obtained.

1. Introduction

Gender-dependent models are more accurate than gender independent ones for speaker indexing, speech recognition, multimedia annotation, speaker adaptation, and speaker recognition. Gender classification can be applied to annotating multimedia archives or to complement video analysis and browsing. Also, gender information can improve speaker clustering, speaker diarization, and voice synthesis. Gender is also an important factor for smart human-computer interaction, social robots, interactive TV, and designing programming environments. Gender classification is an open problem in biometrics. Finally, there is a relationship between spectral cues as well as temporal prosodic ones and the speaker's gender from a phycological point of view.

For the last three decades, a lot of work has been done in gender classification. However, it is still a challenging problem that triggers many researchers. The challenges mainly derive from the fact that gender information is time-invariant, phoneme-independent, and identity-independent for speakers of the same gender [15]. In [8], it is concluded that a 20% error rate in gender classification is common. In [17], it is reported that most of the gender classifications, that are tested for clean speech, demonstrate an accuracy below 95%, and in [13] it is conveyed that automatic gender detection cannot be 100% correct. In [8], it is proven that discriminating between speech and music is easier than separating male from female speakers.

Proceeding toward gender classification in emotional speech, the inherited difficulties of analyzing emotional speech add to the complexity of the problem. There is a lack of comprehensive theoretic reasoning from a phycological point of view [10], because the emotional states are related to a number of human factors, such as mentality and personality, which are hard to model. Previous studies have proven that different genders convey emotions differently. For example, pitch information, which is commonly used, varies over time, since it is driven by an individual's emotional state [5]. It is also influenced by the speakers' need to adapt to the context [5]. In [13], it is proven that if pitch is used as criterion for gender classification, when processing emotional speech instead of spontaneous speech, accuracy is deteriorated by 21%.

Concerning previous work on non-emotional speech, the system proposed by Zeng et al. [17] is based on Gaussian mixture models (GMMs) of pitch and spectral perceptual linear predictive coefficients and its accuracy ranges from 95% to 98.3%. A system using GMMs is proposed by Tzanetakis and Cook that achieves 74% accuracy [11]. Concerning emotional speech, it is true that when gender information is available, emotional speech classification is significantly improved [12]. When emotional speech is processed for gender classification, an accuracy of 94.65% is reported in the Berlin database of Emotional Speech (BDES) by

Xiao et al. [16]. Vogt and Andrè [13] applied gender detection prior to gender-specific emotion classifiers. The BDES and the SmartKom mobile database were used. In the first case, the best reported accuracy equals 90.26%, whereas in the second case it equals 91.85%.

The paper contributions are two-fold. First, novel features are investigated, a significant number of which is finally retained after feature selection. Second, SVM classifiers with different kernels are assessed on two databases of emotional speech. In Section 2, the databases that are used are described and in Section 3 feature extraction and selection are detailed. Experimental results are reported in Section 4. Discussion is made and conclusions are drawn in Sections 5 and 6, respectively.

2 Databases

Two databases are used: the BDES [1] and the Danish Emotional Speech (DES) [7]. In the BDES, 10 actors (5 male and 5 female) simulate 7 emotions: anger, fear, joy, sadness, disgust, boredom, and neutral. A total of 10 German utterances, 5 short utterances and 5 longer ones are recorded. The database comprises of approximately 30 minutes of speech. After evaluation by human subjects, 535 utterances are retained. 233 of 535 are uttered by male speakers, whereas the remaining 302 are uttered by female speakers. The DES database includes utterances expressed by 2 professional actors and 2 actresses in 5 different emotional states: anger, happiness, neutral, sadness, and surprise. The utterances correspond to isolated words, sentences, and paragraphs. The database comprises approximately 30 minutes of speech. Overall, 1160 utterances have been used that are equally split into 580 utterances uttered by male speakers and another 580 utterances uttered by female ones.

3 Feature extraction and selection

The first contribution of the paper is in the selection of a 15-dimensional feature subset extracted from a large set of 1379 features. This procedure has proven to be successful in [13]. Using a small feature set has several advantages, since such sets require less memory, imply less computations, and their statistical modeling is more accurate. Employing large feature sets increases the possibility of including features with less discriminating power. This is experimentally verified by applying all 1379 features on the BDES, where an accuracy of 86.23% is obtained.

The initial feature set includes 2 subsets: features commonly used in emotional speech classification literature [12], and features that are investigated for the first time. The features fall into the following categories: pitch-related, energy-related, spectral, autocorrelation-related, Fujisaki-related [9], jitter- and shimmer- related [4], and features derived from the sound description toolbox [3]. The sound description toolbox extracts the following features: MPEG-7 AudioPower, MPEG-7 AudioFundamentalFrequency, MPEG-7 AudioSpectrumSpread, MPEG-7 AudioSpectrumFlatness, MPEG-7 LogAttackTime, MPEG-7 TemporalCentroid, MPEG-7 AudioSpectrumRolloff frequency, mel-frequency cepstral coefficients (MFCCs), total loudness, and specific loudness sensation (SLS) [3]. The first- and second-order differences of the aforementioned features are computed to capture the feature temporal evolution. All features are subject to min-max normalization, since several features may have a different scale. Min-max normalization preserves all original relationships and does not introduce any bias.

Two classes are considered: emotional speech uttered by a male subject and emotional speech uttered by a female subject. We presume that the class-dependent feature distributions are such that the mean vectors of the two classes are easily discriminated. The goal is to find a subset $F_i(D)$ of dimension $D = 15$. The selection criterion is the Fisher ratio $J = \mathrm{tr}(\boldsymbol{S}_w^{-1}\boldsymbol{S}_b)$, where $\mathrm{tr}(\cdot)$ stands for the matrix trace operator, $\boldsymbol{S}_w$ is within-class scatter matrix, and $\boldsymbol{S}_b$ is the between-class scatter matrix. Feature selection goal is to find $F_i(D)$ such that $J(F_i(D)) \geq J(F_j(D)), \quad j \in \{1, \ldots, q(D)\}$, where $q(D)$ is the number of distinguishable subsets containing D elements. The branch and bound search strategy using depth-first search and backtracking is employed, since its performance is near optimal [6]. The search process is accomplished by means of a tree structure, consisting of 1379-15+1=1365 levels. A level is comprised by a number of nodes and each node corresponds to a feature subset. At the highest level, there is only 1 node including the full feature set. At the lowest level, there are nodes containing 15 features. The search process takes place from the highest level by systematically traversing all levels, until the lowest level is reached. The traversing algorithm uses depth first search with a backtracking mechanism. So, if the best performance found so far is J_1, then branches whose performance is less than J_1 are skipped.

A portion of 20% of the audio recordings in DBES is retained for feature selection. Starting from the most effective feature, the 15 best selected features for the BDES are: the minimum of the 15th MFCC, the mean pitch, the mean of the 12th MFCC, the interquartile range of phrase commands, the minimum of the 11th MFCC, the mean of the 7th MFCC, the median of the

21th MFCC, the mean of the 8th MFCC, the variance of the 17th MFCC, the interquartile range of the 16th MFCC, the interquartile range of the 22th MFCC, the variance of the 1st formant, the mean value of the 1st formant, the minimum of 18th MFCC, the mean variance of the 2nd MFCC. The same approach was repeated for the DES database. The selected 15 best features are: the median of pitch, the variance of the 6th SLS second-order differences, the 90th percentile of the 4th formant, the interquartile range of fundamental frequency first-order difference, the 90th percentile of the 3rd coefficient of AudioSpectrumFlatness first-order difference, the 90th percentile of the 4th coefficient of AudioSpectrumFlatness first-order difference, the interquartile range of energy values within the rising slopes of energy contours, the skewness of the 15th MFCC, the minimum of the 8th MFCC, the minimum of the 4th MFCC, the maximum of the 19th MFCC, the variance of the 21st MFCC, the skewness of the 24th MFCC, the variance of AudioSpectrumRolloff frequency second-order differences, and the variance of the 3rd MFCC. MFCCs are widely selected in both databases. MFCCs are among the most discriminating parameters for gender classification [8]. Pitch is potentially higher for females [5], while formants' efficiency has been demonstrated in [14]. Energy is also commonly used [13]. The fact that the novel features, proposed here, are retained confirms their discriminating ability for gender classification.

4 Experimental Procedure

SVMs are supervised learning methods that avoid overfitting by finding the maximum-margin hyperplane, which achieves the best class separation. Accordingly, they exhibit good generalization performance. SVMs minimize structural risk. SVMs condense all the information contained in the training set relevant to classification in the support vectors, which reduces the size of the training set [14]. This is ideal for the case under consideration, since emotional speech databases are sparse [10]. Quadratic programming is used for solving the optimization problem during training. SVMs with 5 different kernels K, are used, namely: 1. Gaussian radial basis function (RBF) SVM, (SVM1) with $K(\mathbf{x}_{t_i}, \mathbf{x}_{t_j}) = \exp\left\{-\gamma||\mathbf{x}_{t_i} - \mathbf{x}_{t_j}||^2\right\}$; 2. Multilayer perceptron SVM (SVM2) with $K(\mathbf{x}_{t_i}, \mathbf{x}_{t_j}) = S\left(\mathbf{x}_{t_i}^T \mathbf{x}_{t_j} - 1\right)$, where $S(\cdot)$ is a sigmoid function; 3. Quadratic SVM (SVM3) with $K(\mathbf{x}_{t_i}, \mathbf{x}_{t_j}) = \left(\mathbf{x}_{t_i}^T \mathbf{x}_{t_j} + n\right)^q$ with $q = 2, n = 1$; 4. Linear SVM (SVM4) with the same kernel as for quadratic, but with $q = 1$; 5. Polynomial SVM (SVM5) with the same kernel as for SVM3, but with $n = 0$.

The classification results are calculated using a 10-fold cross-validation evaluation. The database to be evaluated is randomly partitioned so that 10% of the utterances are used for testing and 90% are used for training. The process is iterated with different random partitions and the results are averaged. SVM input vectors for each database and each iteration are those described in Section 3. The mean best classification accuracy for all 5 SVM alternatives, along with the corresponding standard deviation (std) is shown in Table 1.

Table 1. SVM alternatives for emotional speech classification accuracy

SVM alternatives	BDES accuracy		DES accuracy	
	mean (%)	std (%)	mean (%)	std (%)
SVM1, γ = 2.4	100	0	99.40	0.09
SVM2	94.95	1.05	93.79	0.96
SVM3	97.38	0.39	98.80	0.43
SVM4	99.44	0.36	98.71	0.18
SVM5, q = 3	99.07	0.41	99.00	0.88

The overall best classification accuracy is achieved for SVM1. The way γ affects accuracy for both databases can be seen in Figure 1. Small γ values yield poor accuracy for both databases. Accuracy for the DES database seems to reach a maximum with a faster rate than the BDES. As γ increases, the accuracy tends to be asymptotically perfect.

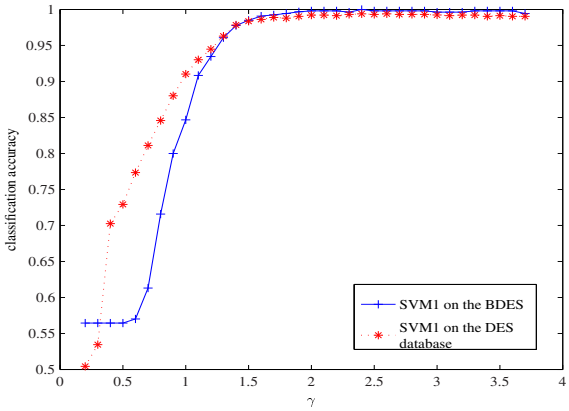

Figure 1. Mean classification accuracy attained by SVM1 for the BDES and the DES databases for various γ values.

5 Discussion

An effort is made to test if the accuracy obtained by SVM1 between the databases is statistically significant.

At 95% confidence level, the t-test value is $1.20\ 10^{-14}$, indicating that SVM1 accuracy differs significantly between the databases. This may be attributed to 2 parameters. First, in [2] it is stated that short-term features have great variability for males and females as is observed for the DES database. Second, a language-dependency is implied.

To allow for a fair comparison with previous approaches, works [13] and [16] are considered, since they both employ the BDES. Concerning [13], an effort is made to compute a multitude of features and subsequently select the best of them, as is true also for our approach. 1289 features are calculated. The feature selection algorithm is a best-first search of a Naive Bayes classifier, which retains 20 features. MFCCs are among the selected features, as happens in our approach, too. The selected classifier for gender classification is Naive Bayes and the best reported accuracy is 90.26%. That is, our approach, when compared to [13] yields a 9.74% improvement in classification accuracy. Concerning [16], 68 features are calculated and 15 are retained, many of which are pitch- and formant-related, as in our approach, too. 15 features and 10-fold cross-validation are employed, as here. A neural network with sequential forward feature selection is applied, which yields a classification accuracy of 94.5%. That is, our approach, when compared to [16] exhibits a 5.35% improvement.

6 Conclusions

In this paper, the problem of emotional speech gender classification is considered. A large feature set comprising of 1379 features is extracted and 15 features are retained through a branch and bound feature selection algorithm. SVMs are applied as classifiers to two databases: the BDES and the DES database. Gaussian RBF SVMs are found to exhibit the best performance, when compared to SVMs with alternative kernels. A perfect classification accuracy is reported, outperforming state-of-the-art techniques.

Acknowledgement

This research project (PENED) is co-financed by E.U.-European Social Fund (75%) and the Greek Ministry of Development-GSRT (25%).

References

[1] Berlin database of emotional speech on-line. http://pascal.kgw.tu-berlin.de/emodb/index-1024.html.

[2] S. M. R. Azghadi, M. R. Bonyadi, and H. Shahhosseini. Gender classification based on feedforward backpropagation neural networks. In *Proc. 4th Int. Conf. Artificial Intelligence Applications and Innovations*, volume 247, pages 299–304. Athens, Greece, September 2007.

[3] E. Benetos, M. Kotti, and C. Kotropoulos. Large scale musical instrument identification. In *Proc. 4th Sound and Music Computing Conference*, July 2007.

[4] P. Boersma. Stemmen meten met Praat. *Stem-, Spraak-en Taalpathologie*, 12:237–251, 2004.

[5] P. Castellano, S. Slomka89, and P. Barger. Gender gates for telephone-based automatic speaker recognition. *Digital Signal Processing*, 7(2):65–79, 1997.

[6] F. V. der Heijden, R. P. W. Duin, D. de Ridder, and D. M. J. Tax. *Classification, Parameter Estimation and State Estimation: An Engineering Approach Using MATLAB*. Wiley, London, U.K., 2004.

[7] I. S. Engberg and A. V. Hansen. Documentation of the danish emotional speech database (DES). Internal aau report, Center for Person, Kommunikation, Aalborg Univ., Denmark, 1996.

[8] H. Harb and L. Chen. A general audio classifier based on human perception motivated model. *Multimedia Tools and Applications*, 34(3):375–395, 2007.

[9] H. Mixdorff. A novel approach to the fully automatic extraction of Fujisaki model parameters. In *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, volume 3, pages 1281–1284. Istanbul, Turkey, June 2000.

[10] J. Rong, Y.-P. P. Chen, M. Chowdhury, and L. Gang;. Acoustic features extraction for emotion recognition. In *Proc. 6th Int. Conf. Computer and Information Science*, pages 419 – 424. Melbourne, Australia, July 2007.

[11] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Trans. Audio, Speech, and Language Processing*, 10(5):293– 302, 2002.

[12] D. Ververidis and C. Kotropoulos. Automatic speech classification to five emotional states based on gender information. In *Proc. of 12th European Signal Processing Conf.*, pages 341–344. Vienna, Austria, September 2004.

[13] T. Vogt and E. Andrè. Improving automatic emotion recognition from speech via gender differentiation. In *Proc. Language Resources and Evaluation Conf.*, page 11231126. Genoa, Italy, May 2006.

[14] L. Walavalkar, M. Yeasin, A. Narasimhamurthy, and R. Sharma. Support vector learning for gender classification using audio and visual cues. *Int. J. Pattern Recognition and Artificial Intelligence*, 17(3):417–440, 2003.

[15] K. Wu and D. G. Childers. Gender recognition from speech. Part I: Coarse analysis. *J. Acoust. Soc. of Am.*, 90(4):1828–1840, 1991.

[16] Z. Xiao, E. Dellandréa, W. Dou, and L. Chen. Hierarchical classification of emotional speech. Technical Report RR-LIRIS-2007-06, LIRIS UMR 5205 CNRS, 2007.

[17] Y. Zeng, Z. Wu, T. Falk, and W. Y. Chan. 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. Dalian, China, 2006.