Practical Considerations for Real-Time Implementation of Speech-Based Gender Detection

Erik Scheme[1], Eduardo Castillo-Guerra[2],
Kevin Englehart[1,2], and Arvind Kizhanatham[3]

[1] Institute of Bimedical Engineering, University of New Brunswick,
P.O. Box 4400, Fredericton, NB, Canada, E3B 5A3
[2] Dept. of Electrical and Computer Engineering, University of New Brunswick,
P.O. Box 4400, Fredericton, NB, Canada, E3B 5A3
[3] Diaphonics Inc., 1310 Hollis Street, Halifax, Nova Scotia, Canada B3J 3P3
{escheme, eduardo.castillo, kengleha}@unb.ca

Abstract. This paper describes a detailed analysis and implementation of a robust gender detector for audio stream applications. The implementation, based on melcepstral features and a Gaussian mixture model classifier, is designed to maximize gender classification performance in continuous speech. The described detector outperforms other reported systems based on statistically significant numbers of gender verifications (2136 unique speakers) obtained from the FISHER speech corpus. The system yields high accuracies for long and short utterances while a confidence figure of merit score for the decision ensures reliability in continuous audio streams.

Keywords: Gender detection, GMM classification, audio streaming.

1 Introduction

The importance of accurate speech-based gender detection is rapidly increasing with the emergence of technologies which exploit gender information to enhance performance. Currently, gender identification is used in security-related applications such as gender mining large volumes of audio recordings, automatic speech monitoring, automatic data labeling and multimedia indexing. Other applications use gender information to train more effective models for speech recognition or speaker identification and verification. Some commercially oriented applications use gender detection for closed captioning and gender-oriented advertisement in audio driven applications. The emergence of these new applications imposes demanding requirements on gender detection system, which may include one or all of the following: Real-time audio stream processing; high confidence for the decision and analysis of a limited amount of useful speech.

Previous investigations in gender identification have proposed a variety of features and classification techniques. Feature extraction is often performed using gender related characteristics of speech such as pitch [1],[3], formant and harmonic structure

J.F. Martínez-Trinidad et al. (Eds.): CIARP 2006, LNCS 4225, pp. 426 – 436, 2006.

[3],[4]. Other approaches rely on spectral features such as Mel-Frequency Cepstral or Spectral Coefficients (MFCC or MFSC) [2],[5], Linear Prediction Coefficients [6], Reflection Coefficients [6] and Log area Ratio Coefficients [5]. Classification techniques use Hidden Markov Models (HMM) [1],[4], Gaussian Mixture Models (GMM)[5],[7],[8] or Neural Networks [2]. Multi-expert approaches have also been developed combining classification techniques [2],[5] .

Despite the abundance of research literature on gender detection, little is focused on its implementation for audio streaming applications and few papers provide practical considerations for performance optimization in real world scenarios.

Harb & Chen [2] have described a general audio classifier for content-based multimedia indexing in continuous speech. They used the mean and variance of 20 MFSCs taken from one second windows to train a collection of eight neural network classifiers based on speech coded with different techniques. They obtained a good dimensionality reduction of features assuming a linear relationship between MFSCs across frames in each one-second segment with results. They performed continuous gender identification with no preprocessing of the incoming signal, so, gender decisions can be made in segments composed solely of silence segments. The gender decisions are also made based upon average MFSCs across one second segments allowing the estimation to possibly include speech from both genders.

This paper describes a detailed analysis of a robust gender detector that addresses some of the limitations observed in previous reports, keeping the classification technique simple to facilitate the implementation. The detector herein is based on a pattern recognition approach where the speech is processed to obtain a representation of the most relevant information for gender identification. The system uses a GMM classifier approach with preprocessed speech, normalized features and provides a decision with a confidence figure of merit (CFM) for each analyzed segment. The performance of two features extraction techniques (MFCC and MFSC [2]) is studied. Different aspects of the GMM are also optimized and practical issues are considered for the real-time implementation of the resulting identification system.

2 Gender Detector

Signal Preprocessing: Audio streaming applications require that a decision be made within a constrained schedule, regulated by specific performance goals. This is typically achieved by analyzing short segments. However, the composition of each processed segment can vary drastically, as well as the amount of noise that is present. Therefore in order to obtain data-independent performance, silent frames are removed using speech activity detection.

The algorithm used herein is based on a combination of zero-crossing (ZC), autocorrelation and energy analysis (EN) of the speech. The ZC analysis discards those frames that the number of crossings is outside a typical range observed in male and female speech (corresponding to a pitch range from 60-400Hz). The computed ZC is normalized by the number of samples corresponding to the pre-selected analysis window (32ms and 8ms increment) according to the sampling frequency of the utterance. The segments that meet the ZC criteria are submitted to autocorrelation analysis where further periodicity of the processed speech is analyzed. A threshold of 0.15 is applied to the normalized correlation values. The segments meeting both previous analyses are then processed with the EN, while the others are used to estimate an adaptive energy threshold. The EN is then performed using two thresholds: an adaptive and absolute threshold. The adaptive threshold is used to discard those frames with energy below the threshold estimated based on the estimated noise segments which should provide an estimation of the utterance noise. The absolute threshold is a fix maximum and minimum hard thresholds were obtained from average telephone speech levels. These thresholds were applied to avoid incorrect estimation of the adaptive threshold due to overly noisy data or other extraneous data conditions. The incoming data stream is continuously preprocessed until a buffer of predefined size is filled with clean data.

Feature Extraction: The feature extraction was tailored to specifically maximize performance for gender detection and not for speech recognition or speaker identification[1]. Consequently, channel compensation and speech normalization techniques were specifically chosen to avoid distortion of gender information. In the speech domain, the mean was subtracted and variance normalization was implemented. The MFCCs were extracted using Hamming windows of 32ms and 8ms increments. In the feature domain, Cepstral mean subtraction[2] and low-pass filtering were implemented. The filtering was used to remove low-amplitude high-frequency content of the spectrum which is highly susceptible to noise and typically consists of unvoiced fricatives containing little information about gender. Variance normalization was not used because it was found to warp the spectral magnitude which decreases observable differences between genders in the extracted features. RASTA [7] technique was found to also decrease the gender discrimination power of the extracted features.

Two variants of Mel frequency coefficients, MFSC and MFCC, were studied. MFSC places emphasis on spectral differences in the mid and high frequencies while MFCC emphasizes differences in the lower spectral content [9]. In each case, 26 Mel filters were used with 19 coefficients and deltas. This effectively results in a low-pass filtering of the framed data in the feature domain at 3 KHz.

Classifier: Classification was performed by selecting the maximum of the log likelihood produced by two gender-dependent GMMs (λ_{Male} and $\lambda_{\mathrm{Female}}$). The expectation maximization algorithm was used to fit the Gaussians to gender-dependent data obtained from the Fisher database [10]. Data from 312 unique speakers of each gender were used to train gender specific models. The amount of training speech per speaker was varied in order to study its effect on classification performance. The optimal number of Gaussian components was also studied in terms of performance and identification speed.

[1] The tailoring of the feature extraction was achieved by empirical optimization of classification performance using the Fisher speech corpus.

[2] Consist on subtracting the mean of each MFCC coefficient for the collection of frames.

Confidence Figure of Merit (CFM) and Decision: A CFM was added to provide feedback about lower quality segments and segments that contain varying levels of speech from both genders. The CFM was estimated by mapping the boundaries of the difference between the mean-log likelihood for each gender model ($\Delta\lambda = \lambda_{\text{Male}} - \lambda_{\text{Female}}$) into the interval [0,1] when varying the composition of testing segments between male and female. In this case, the threshold between genders is mapped into the center of the interval. The effect was observed for 10,000 verifications using 100 unique speakers from each gender. All combinations of the test segments were used to obtain the performance of the system when presented with segments containing both genders (expressed in percentage of speech that is female in Fig 1).

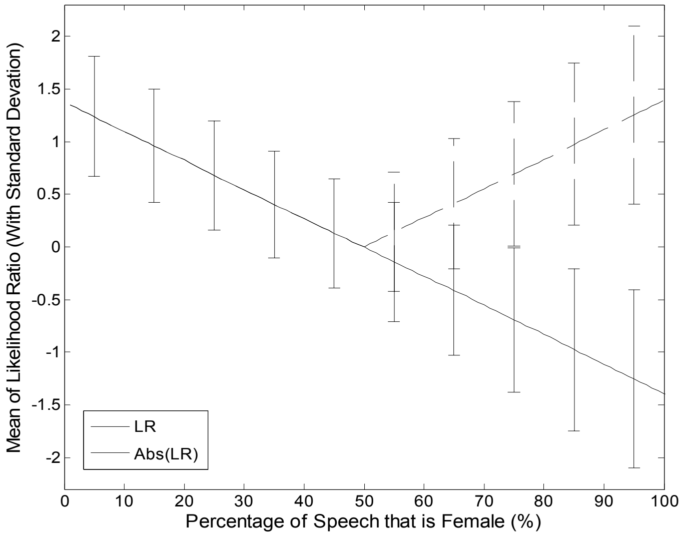

Fig. 1. Likelihood ratio when the composition of the testing segment varies from 0-100% female speech. Dashed lines indicate absolute value.

A positive differential of the log likelihood ratios is shown when the composition consists of more male than female speech indicating that the male model is predominant. However when the speech becomes predominantly female, the differential log likelihood ratio turns negative. A definite trend away from zero exists as a more biased gender composition (greater than 50% male or female) is introduced. This reveals that the system's response is linearly proportional to the constitution of the testing segment and the contribution provided by each gender feature is equally balanced. Mixed gender utterances result in a lower CFM for the decision because their log likelihood ratios tend to be closer to 0. The application controller can therefore monitor for segments with low CFM, and perform further analysis. The gender models used in this section were trained with 30s of speech from 312 unique speakers of each gender and tested with 15s segments extracted from varying combinations of 100 pairs of male-female speakers.

3 Experiments

Three main experiments were conducted throughout this investigation. The first compares the use of MFCC and MFSC features for capturing gender information. Their impact on the performance of models created with variable amounts of speech per speaker (SPS) and variable length of the testing segments (VLTS) was also considered. Pairs of gender models were trained with 60s, 30s, 15s, 10s, 5s and 3s of speech obtained from unique speakers. The models were validated using testing utterances of seven different lengths in order to observe the effect of varying training and testing data length on overall performance.

The second experiment considered the effect of a varying number of GMM components on the performance of gender models created with 30s of SPS and VLTS. Models were trained with 64, 128, 256, 512 and 1024 Gaussians and tested using test utterances of several different lengths.

The last experiment tests the performance of the best performing models from previous sections (MFCC features, 30s SPS, 512 Gaussians) with a continuous stream of speech. The audio stream was composed of alternating male/female speech segments of random-size (between 20 and 30s) from 200 unique speakers taken from the National Institute of Standard Technology (NIST) evaluation data, 2005. Time labels of gender transitions were maintained and used to determine the performance of the system. Trials were completed using 1, 3 and 5 second segments.

In all tests, the models were trained with speech from 312 unique gender-specific speakers obtained from the Fisher database [10]. All utterances were manually verified prior to use to avoid mislabeled and cross-gender cross-talk content. The testing speech for experiments 1 and 2 was taken from 2136 unique speakers of NIST evaluation data, 2005.

4 Results

MFCC-MFSC Performance: The performance of the gender detection system, when extracting MFCC and MFSC features, is shown in Tables 1 and 2. The accuracy obtained when using each feature set is shown for varying lengths of training and testing data for each gender and overall.

A comparison between both tables show that MFCCs outperform MFSCs, with the difference approached 3%. These results contradict those reported in [2], however, the classification approach used herein also differs from that used by Harb & Chen. It can be observed from Table 1 that the overall performance decreases with diminishing amounts of training data. However, the performance for individual gender does not show the same trend for both genders. This could be originated because providing less training speech the model captures more non-stationary characteristics of the waveform disguising the boundaries between genders. It is appreciated that female models perform more consistently across changes in training set length than males. This can be caused due to female gender information is more readily captured by the MFCC features and requires less data amounts to perform compared to male gender.

Table 1. System Performance with MFCC Coefficients

		Gender	Speech Per Speaker Used to Train Gender (Seconds)					
			60	**30**	**15**	**10**	**5**	3
Length of the Testing Segments (Seconds)	**60**	**M**	**96.01**	**96.01**	**95.58**	**95.69**	**95.36**	95.03
		F	**98.51**	**98.59**	**98.68**	**98.51**	**98.18**	97.93
		Both	97.43	97.47	97.33	97.28	96.95	**96.67**
	30	**M**	**95.79**	**96.12**	**95.90**	**95.90**	**95.79**	95.47
		F	**98.68**	**98.43**	**98.43**	**98.10**	**97.44**	97.44
		Both	97.43	97.42	97.33	97.14	96.72	**96.58**
	15	**M**	**95.36**	**95.68**	**94.50**	**95.47**	**94.50**	94.28
		F	**98.35**	**98.18**	**98.59**	**97.51**	**97.68**	97.60
		Both	97.05	97.09	96.82	96.63	96.30	**96.16**
	5	**M**	**93.96**	**93.42**	**93.42**	**91.91**	**92.56**	91.15
		F	**96.94**	**97.44**	**96.86**	**97.44**	**95.62**	96.44
		Both	95.65	95.69	95.37	95.04	94.29	**94.15**
	3	**M**	**93.96**	**90.83**	**90.94**	**90.72**	**89.32**	89.21
		F	**94.79**	**97.02**	**96.69**	**96.61**	**96.53**	95.12
		Both	94.43	94.33	94.19	94.05	93.40	**92.56**
	1	**M**	**87.91**	**90.06**	**88.66**	**86.93**	**88.76**	86.61
		F	**95.21**	**92.98**	**94.30**	**94.63**	**91.32**	92.40
		Both	**92.04**	**91.71**	**91.85**	**91.29**	**90.22**	**89.89**

The spectral differences between each gender, including pitch, formants and high frequency energy distribution also accentuate this difference, impacting differently the MFCC estimation.

The testing utterance also has a direct impact on the performance of the system, with larger testing segments providing best performances. This is because larger testing segments contain more information. However, it is noticeable that the effect of the length of training data is more profound. This is beneficial for streaming applications because while training can be performed offline with large amounts of data, high performance is desired with shorter test utterances.

Upon further analysis of Table 1, it is appreciated that the performance gap between both genders increases while the length of the testing segments decrease. This is created by a combination between the differences in the descriptive power of the MFCC features for the female gender and the decrement of information provided. This creates a fussier delimitation between both genders, causing overlap among them. Table 1 shows that no substantial gain when more than 15s of speaker data is used for training the models. Therefore, this is a good tradeoff between performance and speech length requirements to build the gender models. Table 2 shows deeper differences in data trends for each gender since MFSC capture different information. However, the overall trend keeps decreasing when the length of the training and testing sets decrease.

Table 2. System Performance with MFSC Coefficients

		Gender	Speech Per Speaker Used to Train Gender (Seconds)					
			60	**30**	**15**	**10**	**5**	3
Length of the Testing Segments (Seconds)	**60**	**M**	**95.79**	**96.12**	**96.87**	**97.09**	**98.17**	98.38
		F	**96.20**	**95.53**	**93.88**	**92.72**	**88.01**	86.52
		Both	95.99	95.83	95.38	94.90	93.09	**92.45**
	30	**M**	**94.82**	**95.79**	**96.44**	**97.20**	**98.06**	98.38
		F	**96.20**	**95.53**	**93.80**	**92.14**	**87.01**	84.78
		Both	95.51	95.66	95.12	94.67	92.54	**91.58**
	15	**M**	**94.82**	**95.47**	**96.53**	**96.98**	**98.17**	98.27
		F	**96.77**	**95.12**	**93.22**	**91.89**	**86.10**	84.53
		Both	95.80	95.30	94.88	94.44	92.14	**91.40**
	5	**M**	**92.56**	**93.20**	**93.96**	**95.04**	**96.44**	96.66
		F	**94.87**	**94.46**	**92.14**	**90.07**	**84.45**	82.55
		Both	93.72	93.83	93.05	92.56	90.45	**89.60**
	3	**M**	**90.40**	**90.51**	**91.91**	**92.23**	**94.28**	94.82
		F	**94.79**	**93.80**	**91.81**	**90.41**	**83.95**	81.64
		Both	92.59	92.15	91.86	91.32	89.19	**88.23**
	1	**M**	**86.50**	**87.58**	**89.01**	**90.36**	**93.20**	93.30
		F	**91.65**	**91.16**	**89.04**	**87.60**	**82.56**	79.92
		Both	**89.07**	**89.36**	**89.02**	**88.98**	**87.88**	**86.61**

Effect of Number of Gaussians on Performance: The performance of models with different number of Gaussians is shown in Table 3. It can be seen that accuracy of decisions increases as the number of Gaussians increases, but the computational load increases likewise. For this reason, the optimal selection of performance (a tradeoff of speed and accuracy) may be system specific. Applications requiring extremely fast processing may forfeit minimal accuracy to achieve speed gains. Offline applications desiring peak accuracies may choose a larger number of Gaussians. Herein, 256 Gaussians were chosen as the most efficient compromise between speed and accuracy. Lower number of Gaussians than 64 followed the same trend as the values shown in the table with the respective decrement in performance.

Audio Stream Evaluation: Fig. 2 shows the performance of the system applied to a pseudo-random streaming audio input. The continuous curve represents the gender composition of each tested segment (where 1 signifies 100% male and -1 signifies 100% female). The stem plot denotes the difference of likelihood ratios (clipped to ±1 for visibility) obtained for the previous segment. It can be seen that the decision made in the segments containing speech from both genders (shown as a change in sign of the continuous curve) produce lower likelihood values. These lower scores will correspond to a decision with low CFM. It may therefore be desirable for an application to further scrutinize, or reject, low CFM for the decisions which tend to indicate cross-gender segments. Given this possibility, it is interesting to note the performance of the system under specific CFM restrictions.

Table 3. System Performance with Variable Number of Gaussians

		Gender	Number of Gaussians*				
			1024	**512**	**256**	**128**	64
Length of the Testing Segments (in Seconds)	**60**	**M**	**97.19**	**97.74**	**95.79**	**94.82**	94.61
		F	**97.35**	**96.03**	**95.20**	**94.79**	94.38
		Both	97.27	96.88	95.50	94.81	**94.49**
	30	**M**	**96.65**	**97.09**	**94.82**	**94.39**	94.18
		F	**97.02**	**96.28**	**94.96**	**94.46**	94.54
		Both	96.83	96.68	94.89	94.43	**94.36**
	15	**M**	**96.11**	**96.66**	**94.93**	**94.39**	94.07
		F	**96.94**	**96.11**	**94.87**	**94.79**	94.13
		Both	96.53	96.38	94.90	94.59	**94.10**
	5	**M**	**94.60**	**95.47**	**92.13**	**91.91**	91.48
		F	**96.36**	**94.79**	**93.38**	**93.55**	92.39
		Both	95.48	95.13	92.75	92.73	**91.93**
	3	**M**	**93.74**	**93.74**	**89.97**	**89.64**	88.89
		F	**96.20**	**94.29**	**93.30**	**92.97**	92.06
		Both	94.97	94.02	91.63	91.31	**90.47**
	1	**M**	**90.28**	**91.36**	**86.61**	**86.27**	86.01
		F	**91.40**	**90.58**	**90.89**	**90.33**	89.73
		Both	**90.84**	**90.97**	**88.75**	**88.30**	**87.87**

* Gender models were trained with 30s of speech from 312 speakers per gender.

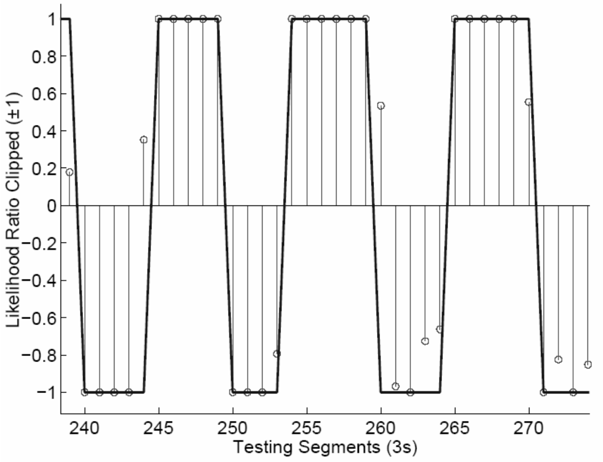

Fig. 2. System performance when applied to an audio stream. Continuous line: composition of the testing segment. Stems: Δλ of segment.

Specifically, Fig. 3 (top) shows the overall performance of the system when a minimum CFM threshold is applied. Fig. 3 (bottom) shows the percentage of total frames that fall into the accepted CFM. Operation of the system without CFM restrictions yields accuracies of 91%, 92.3% and 92.1% on all segments for 1, 3 and 5 second testing segments, respectively. By applying a CFM threshold of 70% -only those segments with CFM over 70% are considered- the system would yield an accuracy of 96.18%, 97.02% and 96.73% (for 1, 3 and 5 second segments), using approximately 80% of the segments.

Computational Complexity: The computational complexity required for the implementation of the gender detector is proportional to the number of Gaussians components used as well as the processing time. The implementation of the algorithms reported was accomplished code generated with Visual Studio 2005 on a 3.2GHz Xeon processor based workstation. The system required 147ms to provide a decision with the most computational demanding setting using 60s of testing speech and 1024 Gaussians. For the setting using 64 Gaussians and 1s of testing speech provided the system required 11ms to provide a decision. In all cases the system performed several times real-time.

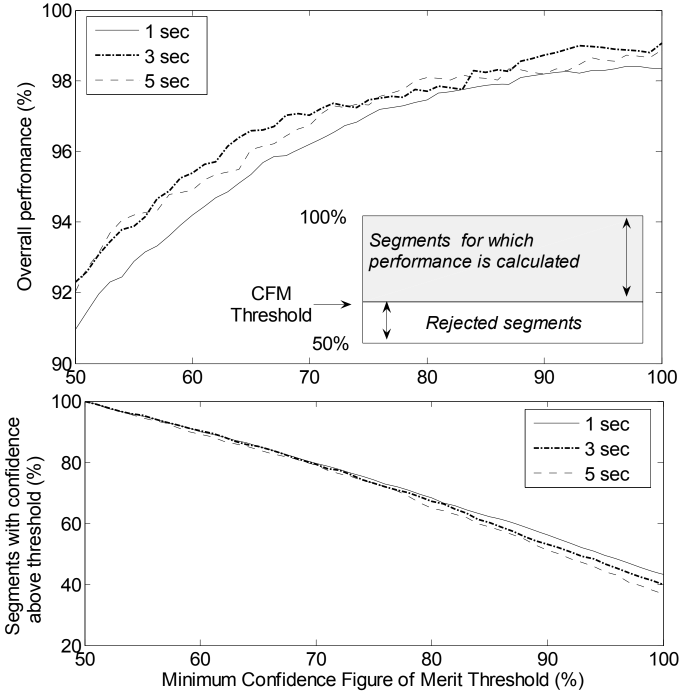

Fig. 3. System performance for testing segments with CFM above threshold (top). Segments meeting minimum CFM threshold (bottom).

5 Conclusion

The experiments performed herein address issues dealing with the implementation of a gender identification system tailored to operate efficiently for audio stream applications. It was observed that the MFCC features that emphasize gender differences observed in the lower part of the spectrum provide more discriminative ability than MFSC features which accentuate the upper spectral band. It was shown that a simple GMM classification approach can be used effectively to provide high classification performance with short-length testing segments performing better or similar to other more complex techniques reported.

The results obtained with the classifier revealed that the amount of speech per speaker used to train the gender models highly influences the performance of the system. Fifteen seconds of speech per speaker was found to be sufficient for training gender models, since longer sequences did not significantly improve performance.

The performance of the system was observed to improve with a greater number of Gaussians, with diminishing returns beyond 512 Gaussians. For streaming audio applications, the optimal number of Gaussians should be chosen to balance accuracy and real-time viability. This research indicated that the system described herein, can reliably perform gender classification at several times real-time in audio streams.

The CFM described enables a measure of quality in the gender decision and improve significantly the performance of the system when a threshold greater than 50% is selected, as showed in experiment 4.3. This metric is very desirable to aid in the detection of cross-gender segments, and to permit higher precision of audio stream applications.

References

1. Parris, E. S., Carey, M. J.: Language Dependent Gender Identification. Acoustics, Speech, and Signal Processing. ICASSP-96 Conference Proceedings, vol. 2. (1996) 685 - 688.
2. Hurb, H., Chen, L.: Gender Identification Using a General Audio Classifier. ICME '03 Proceedings, vol. 2. July (2003) 733-736.
3. Kamran, M., Bruce, I. C.: Robust Formant Tracking for Continuous Speech with Speaker Variability. IEEE Trans. Speech and Audio Proc. Accepted for publication, Jan. 19, (2005).
4. Vergin, R., Farhat A., O'Shaughnessy D.: Robust Gender-dependent Acoustic-phonetic Modelling in Continuous Speech Recognition Based on a New Automatic Male/female Classification. ICSLP-96 Conference Proceedings, vol. 2. October (1996) 1081-1084.
5. Slomka, S., Sridharan, S.: Automatic Gender Identification Optimised for Language Independence. TENCON '97 IEEE Region 10 Annual Conference. Speech and Image Technologies for Computing and Telecommunications Conference Proceedings, vol. 1. December (1997) 685 - 688.
6. Childers, D. G., Ke, W., Bae, K. S., Hicks, D.M.: Automatic Recognition of Gender by Voice. Acoustics, Speech, and Signal Processing. ICASSP-88 Conference Proceedings, vol. 1. (1988) 603-606.

7. Torres-Carrasquillo, P. A., Singer, E., Kohler, M. A., Greene, R. J., Reynolds, D. A., Deller, J. R.: Approaches to Language Identification Using Gaussian Mixture Models and Shifted Delta Cepstral Features, International Conference in Spoken Language. Denver. (2002).
8. Chen, T., Huang, C., Chang, E., Wang, J.: Automatic Accent Identification Using Gaussian Mixture Models. Workshop in Automatic Speech Recognition and Understanding ASRU '01. (2001) 343 – 346.
9. Andrianaki, I., White, P. R.: Modeling of Mel Frequency Features for Non Stationary Noise. Institute of Sound and Vibration Research. University of Southampton. Available: http://dea.brunel.ac.uk/cmsp/ Projnoise2003/Presentation25052004Ioannis.ppt.
10. Fisher English Training Speech Part 1, Linguistic Data Consortium, LDC2004S13, 2004.