An Analysis of the Influence of Acoustical Adverse Conditions on Speaker Gender Identification

Tomasz Maka and Piotr Dziurzanski

Faculty of Computer Science and Information Technology
West Pomeranian University of Technology, Szczecin
Zolnierska 49, 71-210 Szczecin, Poland
Email: tmaka@wi.zut.edu.pl

***Abstract*—Speaker gender as a biometric feature plays an important role in numerous voice-based services. In this work we perform an accuracy analysis of a gender recognition system in different acoustical environments (indoor and outdoor auditory scenes). At the evaluation stage, each sentence has been mixed with several types of background noise using various signal-to-noise ratio levels. Then a voiced parts of speech have been extracted and parametrized using features based on filter banks and vocal-tract properties. The obtained feature trajectories have been non-linearly smoothed in order to minimize the influence of adverse conditions on the spoken sentences. The observed accuracy is acceptable for voice-based tasks where the gender information can improve their performance.**

I. Introduction

Many voice-based services use speaker specific information to improve their accuracy and to compensate acquisition conditions while recording the speech. One of the most important properties of a speaker, exploited in such tasks, is gender information. The problem of automatic gender recognition (AGR) is broadly covered in the literature [1], [2], [3]. However, due to many factors connected with a speaker (emotional state, age, health condition, various acquisition environments, etc.) it is still considered as a challenging problem. The typical applications of an AGR include a speaker recognition and verification, annotation of multimedia databases, human-robot/computer interaction and voice-based biometric systems. The efficiency of an AGR is dependent on the type of features used at the parametrization stage. As in many speaker-specific voice analysis systems, the most frequent features are calculated based on vocal tract attributes. The typical features include fundamental frequency and formant properties. Also, features based on a characteristic of glottal waveforms are often used in such systems. Unfortunately, the performance of voice-based systems is strictly connected with an acquisition environment and background noise can deteriorate the final accuracy significantly. Therefore, the dedicated audio features and an optional post-processing stage shall be used in order to improve the robustness of the system in such situations.

In [2], the AGR system based on two sets of hidden Markov models for male and female speakers has been proposed. An additional post-processing stage has been used to normalise the models and reduce bias towards the selected gender. The obtained results for British English give less than 1% error rate. The approach based on a large feature set has been shown in [3]. The set comprising of 1379 low-level features has been exploited in a feature selection process giving 15 features in final set for AGR task. At the classification stage, the authors obtained accuracy up to about 99 per cent with SVM classifier (using Gaussian RBF kernel) for two emotional databases. The study presented in [4] showed that a score-level fusion with an AdaBoost algorithm can outperform the methods using single voice information. Such approach based on MFCC and pitch features fusion leads to identification rate equal to 98.6 % in noiseless conditions. An analysis of cepstral and prosodic features, in the context of AGR, has been evaluated in [5]. The authors show that use of voiced speech, modeling of higher spectral details and use of dynamic features lead to an improved robustness toward mismatched training and test conditions. Even better performance of an AGR system in adverse environment has been observed for cepstral features.

In this work, an analysis of the influence of acoustic environment on an automatic gender recognition has been performed. The next section describes an architecture of the proposed AGR system and its properties. In section III, a set of experiments has been described: voicing detection, selection of the features and classification accuracy in clean and adverse conditions. The last section concludes the paper with a summary and further research directions.

II. Automatic Gender Recognition

Since the speaker-dependent properties of the voice are calculated from voiced parts of speech signal, the proposed system consists of two parts: a voiced part detector and a gender classifier. The architecture of the AGR system is depicted in Fig. 1. From the input signal, a set of feature contours is calculated. The type of features is selected based on their voiced properties. Each of the trajectories is normalized and then thresholded with an adaptively calculated threshold. The thresholded signals are fused together by summing up and computing the signum function. The resulting signal is applied as a voiced mask to the input signal in order to obtain a signal containing only voiced parts of the speech signal.

978-1-4799-3700-4/14/$31.00

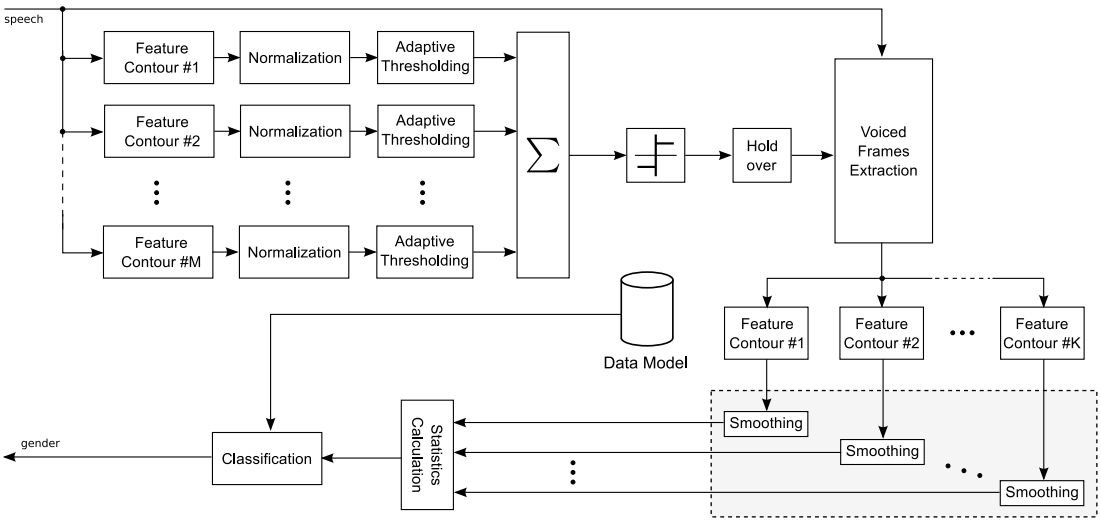

Fig. 1. System for automatic gender recognition based on feature contours analysis and processing.

At this stage we have employed $M = 3$ following feature trajectories to generate a mask for the extracting voiced parts:

- spectral variance (variance of a frame magnitude spectrum),
- AEZR (ratio of energy to zero-crossing rate calculated from the frame autocorrelation signal),
- MOD4HZ (4Hz modulation energy [6]).

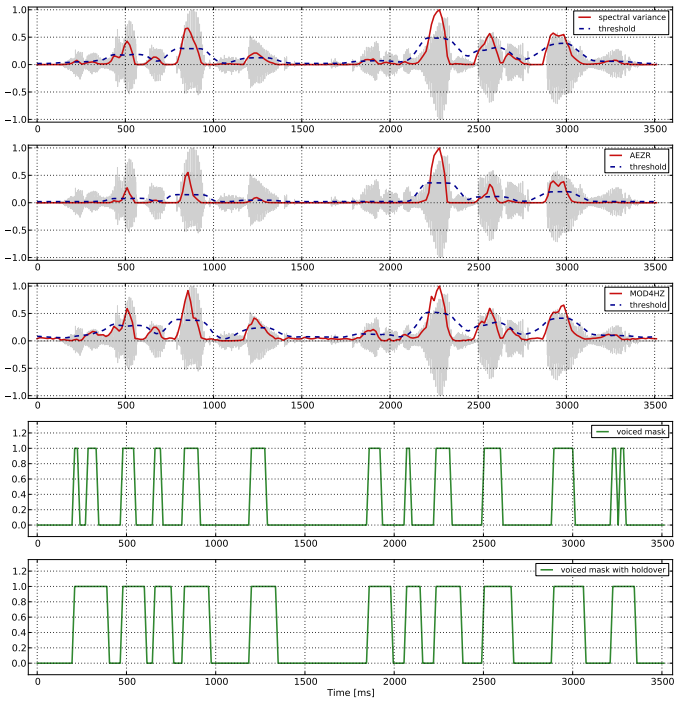

Fig. 2. Voiced parts of speech detection steps (from top to bottom): spectral variance, AEZR and MOD4HZ trajectories with thresholding contours, obtained voiced mask and final mask after holdover operation.

Each of the trajectories is thresholded using an adaptively calculated threshold as described in [7] using parameters $\beta = 1 + \lambda$ and $W = \sqrt{N}$, where λ is an inter-quartile range (iqr) of the trajectory and N is its size. To obtain a final mask for a voiced part, a holdover [8] stage is applied in order to fill up short gaps and elongate quick changes in the mask. The whole process of voicing detection is shown in Fig. 2.

The extracted parts of a signal are further processed in order to create a final feature vector for AGR. For the obtained signal, a set of K feature trajectories is computed (pitch, formants, etc.). In the next step, a nonlinear smoothing operation is executed in order to compensate discontinuities in the signal. As smoothing modules, we have applied a non-linear smoothing system proposed in [9], composed of two blocks: median and linear smoothers. In our case, the median smoother used a window of 5 samples, and the linear smoother was a 3rd order low-pass FIR filter. From the smoothed version of each trajectory, 16 statistics are calculated (including min, max, median, arithmetic mean, skewness, kurtosis, quartiles, etc.) and merged in feature vectors for the classification stage. The classification is performed using the Support Vector Machines classifier [10] with a polynomial kernel and sequential minimal optimization algorithm for training.

III. Experimental Evaluation

A feature discriminatory procedure has been performed using a single sentence spoken by 630 speakers (438 male and 192 female speakers) from the TIMIT database [11]. The details of data used in the experiments are provided in Tab. I. The audio data split for training and testing parts has been chosen as 75/25 and the total length has been equal to about 36 minutes. In order to perform an analysis of the AGR accuracy in different acoustical environments, we have used a set of background noise recordings [12] and mixed them with sentences from the TIMIT database at three signal-to-noise ratio (SNR) levels (-5dB, 0dB and 5dB). The selection of acoustical scenes has been based on different long-term spectral properties in order to analyse their influence on the

speech signal – their long-term average spectra are shown in Fig. 3.

TABLE I
CHARACTERISTICS OF SELECTED AUDIO DATA FROM TIMIT DATABASE

	Training set		Testing Set	
	Items	Length [s]	Items	Length [s]
Male	326	1092	112	381
Female	136	483	56	198
Total	462	1575	168	579

At the speech parametrization stage, the following features have been employed [9]:

1) Mel-Frequency Cepstral Coefficients (MFCC, 12 trajectories),
2) Linear-Frequency Cepstral Coefficients (LFCC, 12 trajectories),
3) Linear Prediction Coefficients (LPC, 12 trajectories),
4) Fundamental frequency estimated with YIN pitch detector [13] (F0, one trajectory),
5) Formants position (FMNT, 3 trajectories),
6) Formants bandwidth (FMNTBW, 3 trajectories).

The total number of input features has been equal to $43 \cdot 16 = 688$, where the number of trajectories was $K = 43$ with 16 statistical descriptors calculated for each trajectory. The selection of the features has been performed using a correlation-based feature subset selection (CFS) with the best first search strategy [14]. As the result of the feature selection process, a feature vector dimensionality has been reduced from 688 to 52 – the final set is shown in Tab. II (where *mad* denotes median absolute deviation and *q1 / q3* are the first and third quartile respectively). Using the obtained features, we have performed a classification for the selected audio data and the obtained accuracy was equal to 99.4%. In the next experiment, we have used speech sentences mixed with four types of background noise at different SNRs to analyse an AGR accuracy in adverse conditions. To compensate the influence of environmental sounds on speech signals, we have performed a classification for raw and smoothed (as marked in Fig. 1) versions of the feature trajectories. The confusion matrices for AGR performance are presented in Tab. III, where the values in brackets denote results for the smoothed trajectories. As it can be observed, for the background sounds with babble noise and numerous events ('*restaurant*', '*shopping centre*'), the recognition performance is worse than in the case of semi-stationary noise without events ('*passing cars*', '*rain*'). The gender classification accuracy is depicted in Fig. 4. It should be noted that classification based on the smoothed versions of feature trajectories gives the same or improved accuracy in comparison with the unprocessed ones. The highest accuracy gain is obtained with '*rain*' and '*shopping centre*' noises where the improvement equals 1.8% for SNR $= -5dB$. For all the analysed cases, the worst classification performance occurs in case of the '*restaurant*' noise for SNR $= -5dB$ and equals 92%.

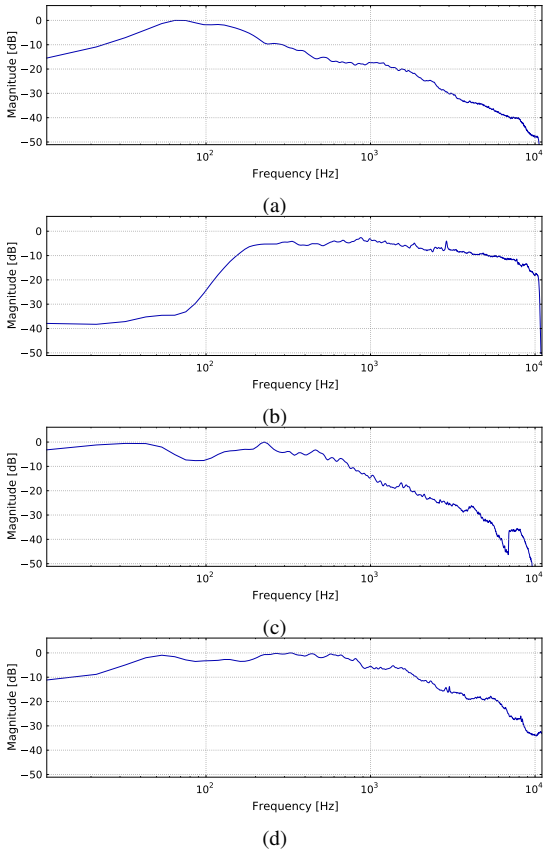

Fig. 3. Long-term average spectra for passing cars (a), rain (b), restaurant (c) and shopping centre (d) background noise.

TABLE II
FEATURES SET OBTAINED IN FEATURE SELECTION PROCESS

Feature	Descriptors		
MFCC-12	MFCC2 / mad, MFCC3 / kurtosis, MFCC5 / iqr, MFCC7 / q1, MFCC9 / max, MFCC12 / iqr	MFCC3 / max, MFCC4 / variance, MFCC7 / min, MFCC8 / mean, MFCC9 / mean,	MFCC3 / midrange, MFCC5 / skewness, MFCC7 / variance, MFCC8 / q1, MFCC11 / max,
LFCC-12	LFCC3 / variance, LFCC5 / midrange, LFCC6 / mad, LFCC9 / max, LFCC9 / variance, LFCC10 / skewness,	LFCC4 / mean, LFCC6 / max, LFCC7 / kurtosis, LFCC9 / midrange, LFCC9 / q3, LFCC12 / q1	LFCC4 / mad, LFCC6 / skewness, LFCC7 / q3, LFCC9 / mean, LFCC10 / mean,
LPC-12	LPC6 / min, LPC10 / q3	LPC6 / max,	LPC7 / skewness,
F0	F0 / min, F0 / kurtosis, F0 / mad,	F0 / variance, F0 / median, F0 / q1,	F0 / skewness, F0 / mode, F0 / iqr
FMNT-3	FMNT1 / q1, FMNT2 / median	FMNT1 / q3,	FMNT2 / midrange,
FMNTBW-3	FMNTBW1 / min,	FMNTBW3 / q3	

IV. CONCLUSION

In this work, a system for automatic gender recognition has been proposed. We have performed analysis of an influence of acoustical adverse conditions on the classification accuracy. For this task, we have prepared a set of acoustic and prosodic features at the feature extraction stage, where the feature

TABLE III
CONFUSION MATRICES FOR GENDER RECOGNITION IN THE PRESENCE OF ENVIRONMENTAL NOISES

Passing cars		
Gender	**MALE**	**FEMALE**
MALE	*5dB / 0dB / -5dB* 111 / 112 / 110 (111 / 112 / 110)	*5dB / 0dB / -5dB* 1 / 0 / 2 (1 / 0 / 2)
FEMALE	*5dB / 0dB / -5dB* 1 / 2 / 3 (1 / 1 / 2)	*5dB / 0dB / -5dB* 55 / 54 / 53 (55 / 55 / 54)
Rain		
Gender	**MALE**	**FEMALE**
MALE	*5dB / 0dB / -5dB* 111 / 110 / 111 (111 / 110 / 111)	*5dB / 0dB / -5dB* 1 / 2 / 1 (1 / 2 / 1)
FEMALE	*5dB / 0dB / -5dB* 0 / 1 / 3 (0 / 0 / 1)	*5dB / 0dB / -5dB* 56 / 55 / 53 (56 / 56 / 55)
Restaurant		
Gender	**MALE**	**FEMALE**
MALE	*5dB / 0dB / -5dB* 110 / 108 / 106 (110 / 108 / 106)	*5dB / 0dB / -5dB* 2 / 4 / 6 (2 / 4 / 6)
FEMALE	*5dB / 0dB / -5dB* 6 / 7 / 8 (4 / 6 / 6)	*5dB / 0dB / -5dB* 50 / 49 / 48 (52 / 50 / 50)
Shopping centre		
Gender	**MALE**	**FEMALE**
MALE	*5dB / 0dB / -5dB* 110 / 108 / 103 (110 / 108 / 105)	*5dB / 0dB / -5dB* 2 / 4 / 9 (2 / 4 / 7)
FEMALE	*5dB / 0dB / -5dB* 3 / 3 / 4 (2 / 2 / 3)	*5dB / 0dB / -5dB* 53 / 53 / 52 (54 / 54 / 53)

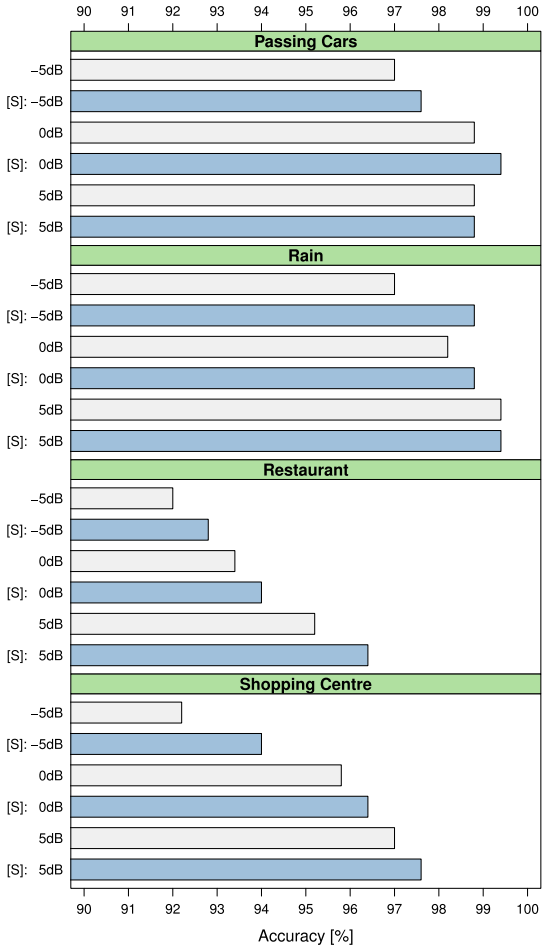

Fig. 4. Gender classification accuracy in different acoustic environments before and after (marked as [S]) the smoothing stage of feature trajectories.

selection process has been exploited. Our experiments show that after introducing a non-linear smoothing stage in the parametrization step, we can gain the accuracy up to nearly 2%. The results indicate that the proposed set of features can be used for speech signals recorded in presence of noisy environment and still maintain the classification ratio over 92%.

Future work will incorporate a further stage to determine the environmental noise, a dedicated post-processing operation chain and different feature sets to compensate a degradation of the input signal.

ACKNOWLEDGMENT

The research work presented in this work was supported by Polish National Science Centre (grant no. N N516 492240).

REFERENCES

[1] H. Harb and L. Chen, *Voice-Based Gender Identification in Multimedia Applications*, Journal of Intelligent Information Systems, 24:2/3, pp. 179–198, 2005.

[2] E. S. Parris and M. J. Carey, *Language Independent Gender Identification*, Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., vol. 2, pp. 685–688, Atlanta, GA, 1996.

[3] M. Kotti and C. Kotropoulos, *Gender Classification In Two Emotional Speech Databases*, Pattern Recognition, 2008. ICPR 2008. 19th International Conference on, pp. 1–4, Tampa, FL, 2008.

[4] M. Ichino and N. Komatsu and W. Jian-Gang and Y. W. Yun, *Speaker gender recognition using score level fusion by AdaBoost*, Control Automation Robotics & Vision (ICARCV), 2010 11th International Conference on, pp. 648–653, Singapore, 2010.

[5] M. Pronobis and M. Magimai.-Doss, *Analysis of F0 and Cepstral Features for Robust Automatic Gender Recognition*, Technical Report, LIDIAP-REPORT-2009-020, 2009.

[6] E. Scheirer and M. Slaney, *Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator*, In Proc. ICASSP-97, Apr 21–24, Munich, Germany, 1997.

[7] T. Maka, *Features of Average Spectral Envelope for Audio Regions Determination*, International Conference on Signals and Electronic Systems - ICSES12, Sep. 19-21, Wroclaw, Poland, 2012.

[8] A. Peinado and J. Segura, *Speech Recognition Over Digital Channels – Robustness and Standards*, John Wiley & Sons, Ltd, 2006.

[9] L. Rabiner, R. Schafer, *Theory and Applications of Digital Speech Processing*, Pearson Higher Education, Inc., 2011.

[10] C. Chang and C. Lin, *LIBSVM : a library for support vector machines*, ACM Transactions on Intelligent Systems and Technology, 2:27:1–27:27, 2011.

[11] J. Garofolo, et al., *TIMIT Acoustic-Phonetic Continuous Speech Corpus*, Linguistic Data Consortium, Philadelphia, 1993.

[12] T. Maka, *Environmental Background Sounds Classification Based on Properties of Feature Contours*, Recent Trends in Applied Artificial Intelligence Lecture Notes in Computer Science Volume 7906, 2013, pp. 602-609, IEA/AIE 2013, Amsterdam, 2013.

[13] A. de Cheveigne and H. Kawahara, *YIN, a fundamental frequency estimator for speech and music*, The Journal of the Acoustical Society of America, 111:1917, 2002.

[14] M. Hall, *Correlation-based Feature Subset Selection for Machine Learning*, PhD thesis, Department of Computer Science, University of Waikato, Waikato, N.Z, 1999.