EFFECTS OF GENDER INFORMATION IN TEXT-INDEPENDENT AND TEXT-DEPENDENT SPEAKER VERIFICATION

Anssi Kanervisto, Ville Vestman, Md Sahidullah, Ville Hautamäki, Tomi Kinnunen

School of Computing, University of Eastern Finland

ABSTRACT

It is well-known that for speaker recognition task, gender-dependent acoustic modeling performs better than gender-independent modeling. The practice is to use the gender ground-truth and to train gender-dependent models. However, such information is not necessarily available, especially if speakers are remotely enrolled. A way to overcome this is to use a gender classification system, which introduces an additional layer of uncertainty. To date, such uncertainty has not been studied. We implement two gender classifier systems and test them with two different corpora and speaker verification systems. We find that estimated gender information can improve speaker verification accuracy over gender-independent methods. Our detailed analysis suggests that gender estimation should have a sufficiently high accuracy to yield improvements in speaker verification performance.

Index Terms— Speaker verification, gender dependent system, gender classification

1. INTRODUCTION

Automatic speaker verification (ASV) [1, 2, 3], the task to verify the identity of a speaker, finds applications in forensics, surveillance and user authentication. Although the modern ASV techniques, such as *Gaussian mixture model – universal background model* (GMM-UBM) [4], and *i-vectors* [5] are relatively robust, most assume explicit knowledge of the speaker's gender. Due to physiological differences of female and males [6], leading to different voice qualities [7], many ASV systems employ gender-dependent UBMs (or other system components). At the enrollment stage, a target speaker model is trained on provided gender information, and a test utterance is scored assuming that gender.

Even if gender-dependent ASV models have usually an edge over fully gender-independent models in terms of recognition accuracy, explicit gender information may not always be available, reliable or meaningful. A user authentication service over a remote channel (such as online banking) might have no face-to-face human supervision at any stage, leading to a risk of enrolling a speaker assuming wrong gender, either purposefully or accidentally. The consequences of this to ASV system performance have not been reported in literature. Further, from an ASV point of view, the biological definition of gender might not be even meaningful: a female with a low fundamental frequency (F_0) or a male with a high F_0 might benefit from using the UBM of the opposite gender based on better match acoustics.

An obvious approach is to use a *gender classification* (GC) system to estimate speaker's gender. In this study, we compare different gender classifiers and their integration strategies with ASV system. In a related study [8], *soft* gender labels improved ASV accuracy for cross-gender trials. We do not consider cross-gender trials but focus on a detailed assessment of the role of gender detection to ASV performance. The work is a part of an ongoing H2020-funded OCTAVE project[1] that develops ASV to physical and logical access control including remote enrollment scenarios. The importance of gender has also been noted by the National Institute of Standards and Technology (NIST) in their on-going 2016 NIST Speaker Recognition Evaluation (SRE) campaign that, in contrast to the earlier SREs, is conducted in a gender-blind manner [9].

Our experiments include both text-independent and text-dependent experiments as well as comparison of soft and hard gender labels. A specific research question that we are unaware of being addressed in earlier studies concerns the impact of gender detection accuracy to the ASV accuracy. Thus, we simulate a gender detector that provides correct gender label according to a certain probability. Our analysis reveals how accurate a gender detector should be in order to produce reliable ASV scores. We recognize usage of gender information in speaker verification/recognition has been well studied in the past (e.g. [8, 10, 11]. The novelty of this paper is studying the effects of inaccurate gender information in AVS.

2. GENDER CLASSIFICATION

2.1. Prior studies

A gender classifier (GC) predicts a speaker's gender based on a provided speech utterance. A summary of selected prior studies on gender classification is given in Table 1. Due to

The paper reflects some results from the OCTAVE Project (#647850), funded by the Research European Agency (REA) of the European Commission, in its framework programme Horizon 2020. The views expressed in this paper are those of the authors and do not engage any official position of the European Commission.

[1] https://www.octave-project.eu/

978-1-5090-4117-6/17/$31.00 ©2017 IEEE

Table 1. *A summary of previous studies on gender classification.*

Reference	Corpus	Number of Speakers	Environment	Method	Accuracy (%)
[12]	In-house	N.A.	Radio recording, telephone, outdoor.	Long-term analysis Mel frequency spectral coefficients (MFSC) with ANN.	91.0
[13]	SRMC	303	Recorded with PDA	MFCC and pitch with GMM	96.7 to 99.7
[14]	Mix of four corpora	460	Clean and noisy	RASTA-PLP with GMM	98.0 (clean) and 95.0 (Noisy-SNR 0 dB).
[15]	aGender	772	Short utterances	Fusion of multiple systems based on MFCC, pitch and others with GMM and SVM.	88.4
[16]	TIMIT and NUST603_2014	630 and N.A.	Microphone, Microphone+telephone +mobile channels	i-vector with PLDA	99.4 to 99.9
[17]	TIMIT and Arabic Database	630 and 71	Microphone	Modified voice contour with SVM	98.3 (TIMIT) 100.0 (Arabic)
[18]	HMIHY	1654	Telephone conversation	F_0 and MFCC statistics with four different classifiers	95.2

the use of different datasets, the accuracies cannot be directly compared, though they all indicate relatively high accuracy.

2.2. Gender classification systems for the current study

We have implemented three different gender classifiers. The two first ones use Gaussian mixture model (GMM) and i-vectors, respectively. Both use 39-dimensional Mel-frequency cepstral coefficient (MFCC) features and discard non-speech parts with energy-based speech activity detector. In addition, combination of MFCCs and F_0 features were tried. The **GMM-based GC** trains two separate GMMs to model male and female features, and a test utterance is classified using a log-likelihood ratio score. Each GMMs uses 128 Gaussians. The **i-vector-based GC** first trains a UBM (256 Gaussians) and a T-matrix (100 dimensions) using data from both genders. The extracted i-vectors are then processed with linear discriminant analysis (LDA) to project them into one dimension taken as the gender score [16]. Our last, **$\mathbf{F_0}$-based** GC system, uses average F_0 of an utterance directly as the gender detection score.

In many prior studies gender detection was treated as an identification task, but we treat it as a detection (2-class) task, by designating arbitrarily either male or female to represent the positive or negative class; this has no effect to our selected evaluation metric, equal error rate (EER).

3. GENDER CLASSIFICATION AND ASV

The direct way to use gender classifier output in ASV system is to take a hard decision from the classifier as the selector of male of female UBM and T-matrix. The other way is to use the soft decision to weight the ASV [10]. In the text-dependent case, the equal error rate among female speakers decreased (4.41% to 2.73%) but for male speakers it increased (1.79% to 1.95%) compared to using oracle hard gender labels [10].

The system in [10] combined the speaker recognition normalized scores S_m and S_f by using the posterior probabilities $\pi(m|X_e)$, $\pi(m|X_t)$, $\pi(f|X_e)$ and $\pi(f|X_t)$, where X_e are the enrollment features, X_t are the test features, $\pi(\cdot)$ is function that returns probabilities of feature belonging to given gender and labels *m* and *f* indicate gender (male and female). These values were combined into final score using.

$$S = \pi(f|X_e)\pi(f|X_t)S_f + \pi(m|X_e)\pi(m|X_t)S_m$$

Experiments in this paper will use the same method, except the scores S_m and S_f are not pre-normalized.

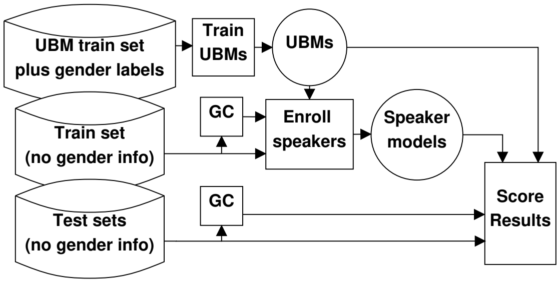

UBM train set plus gender labels
Train UBMs
UBMs
Train set (no gender info)
GC
Enroll speakers
Speaker models
Test sets (no gender info)
GC
Score Results

Fig. 1. *Automatic speaker verification system with gender classification (GC). GC is used to determine which UBM model (male or female) should be used for enrolling and testing speaker utterance. This can be done strictly by selecting correct UBM according to estimated gender or by combining scores obtained from both models.*

4. EXPERIMENTAL SETUP

4.1. Setup for gender classification experiment

Gender classification experiments are conducted on recently released RSR2015 corpus [19] and telephone condition (CC5) of NIST SRE 2010 (SRE10). For the experiments on RSR2015, we used the background set for training the gender models. Two different trial lists were created from the development and evaluation sets. The summary of the corpora for GC experiments are shown in Table 2.

Table 2. *Database description for gender classifier experiments on RSR2015 and NIST SRE 2010.*

Database	Section	Male segments	Female segments
RSR2015	Train	1011	1108
	Dev	9099	9972
	Eval	9059	22220
NIST SRE 2010	Train	3420	2653
	Test	2486	1759

Table 3. *Database description for automatic speaker verification experiments on RSR2015.*

Task	Description	Development		Evaluation	
		Male	Female	Male	Female
a	Target models	1492	1405	1708	1470
	Target trials	8931	8419	10244	8810
	Non-Target trials	437631	387230	573664	422880
b	Target models	50	47	57	49
	Target trials	4443	4205	5116	4404
	Non-Target trials	217718	193431	286496	211392
c	Target models	50	47	57	49
	Target trials	4488	4214	5128	4406
	Non-Target trials	219913	193799	287168	211488

4.2. Setup for automatic speaker verification experiment

For conducting the ASV experiments, we use the same corpora as in the GC experiments. The RSR2015 is used to perform ASV experiments in three different text-dependent and text-independent tasks. These three different protocols range from a pass-phrase situation to a text-independent situation. In protocol (a), system is trained and tested with fixed pass-phrases (one pass-phrase per speaker). In (b), speaker is prompted with one of the possible pass-phrases (multiple pass-phrases per speaker), and in (c), enrollment and test pass-phrases are different (i.e. text-independent). The summary of the database for three different tasks are shown in Table 3. For the experiments with RSR2015, we use the MFCC features with GMM-UBM system similar to [20].

As for the SRE10, we conduct the experiments using the same i-vector-PLDA system as used in [20], but with block-based MFCC features [21].

5. RESULTS

5.1. Performance of stand-alone gender classifier system

We compare the performances of GC systems in terms of equal error rates (EERs), calculated using the BOSARIS toolkit [22]. The results are shown in Table 4 for RSR2015. MFCCs with GMM gives the best performance on both development and evaluation set. Augmenting logF_0 with the MFCC does not help in improving gender classification performance. F_0 thresholding method produces reasonable EERs though is clearly behind our MFCC-based methods, as one may expect.

Similarly, we report the gender detection performance on SRE2010 in Table 5. We also ran experiments with 512 Gaussians and i-vector dimension of 400, lowering the EER of GMM system to 3.68% while having little effect on i-vector system. However, no significant change was obtained with RSR2015 corpus by changing hyper-parameters.

Table 4. *Gender classification performance with RSR2015 set in terms of EER (%).*

Front-end	Backend	Development	Evaluation
MFCC	i-vector	1.36	1.81
MFCC+F_0		1.97	0.93
MFCC	GMM	**1.11**	**0.72**
MFCC+F_0		3.34	1.58
F_0	-	3.75	1.93

Table 5. *Gender classification performance with SRE 2010 corpus in terms of EER (%).*

System	EER
MFCC +i-vector	**4.29**
MFCC +GMM	5.93

5.2. ASV performance with gender classifier system

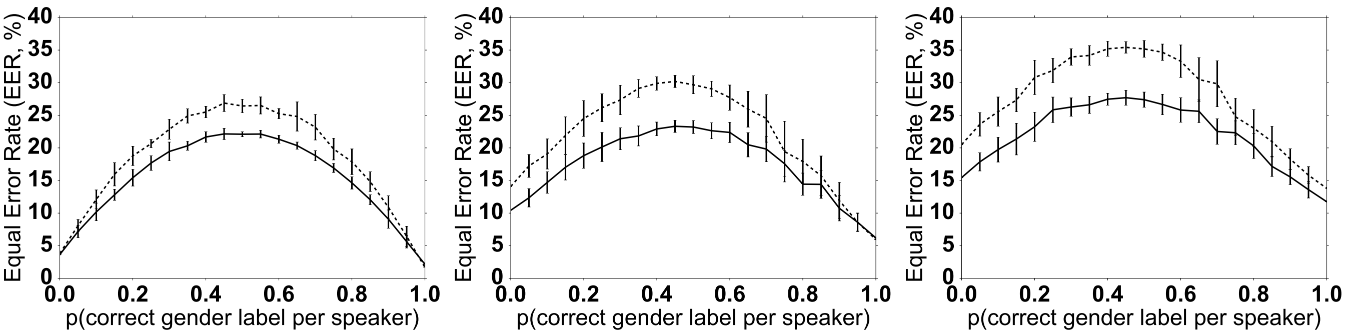
To study effects of gender information in automatic speaker verification, we use the gender information in four different ways. First, we used gender ground-truth provided in the corpus metadata. Second, we did not use any gender information at all and built only one gender-independent ASV system. Finally, in the other two cases, we have used *hard* and *soft* labels provided by the gender classifier. The ASV performance in four different conditions are reported in Tables 6 and 7, correspondingly for RSR2015 and NIST SRE 2010. The numbers in bold face indicate lowest EERs excluding ground-truth 'labels'.

The results indicate that using gender labels improves performance for NIST SRE 2010 where the signals are of telephone channel quality. For RSR2015 corpus, the trained GC performs better but using gender information in ASV provides inconsistent improvements. Summary is found in Table 4.

5.3. Effects of GC accuracy in ASV

To study how much gender classification error can affect speaker verification accuracy, we conduct experiments for worst case scenario with completely wrong labels by the flipping the ground-truth information. The results are shown in Table 8. We observe ≈ 50% relative degradation in EER for both text-dependent and independent scenarios.

To further study the effects of GC accuracy on ASV performance, we performed experiments by assigning simulated gender detector labels to the speakers, as if they were predicted by a GC system. The experiments were conducted by varying the classification error probability of the simulated GC system. The probability of retaining the correct speaker's gender ranged from 0 (i.e., all labels wrong) to 1 (i.e., all labels correct). For each level of p(correct label), EERs were calculated 20 times by randomly picking speakers whose gender labels were flipped based on p(correct label), and the av-



(a) Fixed-pass phrase text-dependent (b) Text-prompted text-dependent (c) Text-prompted text-independent

Fig. 2. *Results of simulating gender classifier accuracy at different levels of correct classification plotted against equal error-rate from speaker verification. Error bars indicate standard error of multiple simulated EERs. Solid line represents male and dashed line represents female speakers.*

Table 6. *Speaker verification results for different tasks on RSR2015 corpus. In each case, the result of the best system not utilizing existing gender labels is emphasized.*

Protocol	Gender	EER(%)			
		Labels	No labels	GC hard	GC soft
a (eval)	F	1.67	2.12	**2.10**	3.01
	M	2.16	**2.21**	2.26	2.34
	All	1.92	**2.17**	2.18	2.68
a (dev)	F	1.86	**2.23**	2.80	3.57
	M	2.83	2.95	**2.92**	3.36
	All	2.35	**2.59**	2.86	3.47
b (eval)	F	5.95	7.37	**6.38**	7.39
	M	6.20	**6.00**	6.27	6.24
	All	6.08	6.69	**6.32**	6.81
b (dev)	F	5.92	7.33	**6.81**	8.06
	M	5.31	**4.58**	5.04	5.33
	All	5.62	5.96	**5.93**	6.70
c (eval)	F	13.80	14.59	**14.23**	14.73
	M	11.74	**10.80**	11.88	11.27
	All	12.77	**12.70**	13.06	13.00
c (dev)	F	12.45	13.40	**13.28**	14.06
	M	9.78	**8.42**	9.96	9.29
	All	11.11	**10.91**	11.62	11.68

erage EERs were computed. These experiments were performed on RSR2015 corpus with different ASV systems.

Figure 2 shows that gender classifier accuracy in speaker verification affects both genders for all three tasks in a similar manner. The EER peaks around the middle of the GC accuracy range and decreases towards both ends $p(\text{correct label}) = 0\%$ and $p(\text{correct label}) = 100\%$. A GC system producing gender decisions close to the chance level (50%) affects the ASV performance most, as one may expect. In this case, the ASV scores are not consistently normalized. But if *all labels are wrong* or *all labels are correct*, accuracy is reasonable.

Table 7. *Results in terms of EER (in %) for SRE10 telephone condition (CC5).*

Gender	Labels	No labels	GC hard	GC soft
F	3.10	4.49	**3.10**	8.75
M	1.98	3.68	**1.99**	3.01
All	3.39	4.03	**3.45**	7.10

Table 8. *Results between using true speaker labels versus using completely wrong (flipped) genders during enrollment and trials.*

Protocol	Gender	EER(%)	
		Labels	Flipped labels
a (eval)	F	1.67	3.81
	M	2.16	3.64
	All	1.92	3.73
b (eval)	F	5.95	14.07
	M	6.20	10.43
	All	6.08	12.25
c (eval)	F	13.80	20.49
	M	11.74	15.44
	All	12.77	17.97
SRE10 (CC5)	F	3.10	17.20
	M	1.98	18.12
	All	3.39	19.49

6. CONCLUSION

We studied ASV performance jointly with a gender classifier. MFCC features with GMM back-end yielded the best results on clean data but i-vector back-end was useful for telephone speech. Further, GC system helps to improve ASV performance when gender information during enrollment and verification is unknown. Further experiments with simulated gender labels reveal the importance of making coherent gender decision, whether *all correct* or *all wrong*. The steep slope of our EER curves close at the endpoints suggests that ASV accuracy might be easily perturbed even by slight degradation in gender detection accuracy. Thus, improving gender detection accuracy in the ASV context involving automatic enrollment or otherwise unsupervised scenarios remains an important practical problem.

7. REFERENCES

[1] D.A. Reynolds and R.C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech and Audio Processing*, vol. 3, pp. 72–83, January 1995.

[2] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: from features to supervectors," *Speech Comm.*, vol. 52, no. 1, pp. 12–40, Jan. 2010.

[3] J.H.L. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal Proc. Mag.*, vol. 32, no. 6, pp. 74–99, 2015.

[4] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1, pp. 19–41, January 2000.

[5] N. Dehak, P.J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788–798, May 2011.

[6] I.R. Titze, "Physiologic and acoustic differences between male and female voices," *The Journal of the Acoustical Society of America*, vol. 85, no. 4, pp. 1699–1707, 1989.

[7] E. Mendoza, N. Valencia, J. Muñoz, and H. Trujillo, "Differences in voice quality between men and women: Use of the long-term average spectrum (LTAS)," *Journal of Voice*, vol. 10, no. 1, pp. 59 – 66, 1996.

[8] M. Senoussaoui, P. Kenny, P. Dumouchel, and N. Dehak, "New cosine similarity scorings to implement gender-independent speaker verification," in *INTERSPEECH*, 2013.

[9] "Speaker recognition evaluation 2016," https://www.nist.gov/itl/iad/mig/speaker-recognition-evaluation-2016, Accessed: 2016-08-15.

[10] T. Stafylakis, P. Kenny, M. J. Alam, and M. Kockmann, "Speaker and channel factors in text-dependent speaker recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 1, pp. 65–78, 2016.

[11] Mohammed Senoussaoui, Patrick Kenny, Niko Brümmer, Edward De Villiers, and Pierre Dumouchel, "Mixture of plda models in i-vector space for gender-independent speaker recognition.," in *INTERSPEECH*, 2011, pp. 25–28.

[12] H. Harb and Liming Chen, "Gender identification using a general audio classifier," in *Multimedia and Expo, 2003. ICME '03. Proceedings. 2003 International Conference on*, 2003, vol. 2, pp. II–733–6 vol.2.

[13] H. Ting, Y. Yingchun, and W. Zhaohui, "Combining MFCC and pitch to enhance the performance of the gender recognition," in *2006 8th international Conference on Signal Processing*, 2006, vol. 1.

[14] Yu-Min Zeng, Zhen-Yang Wu, T. Falk, and Wai-Yip Chan, "Robust gmm based gender classification using pitch and RASTA-PLP parameters of speech," in *2006 International Conference on Machine Learning and Cybernetics*, 2006, pp. 3376–3379.

[15] M. Li, K.J. Han, and S. Narayanan, "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion," *Computer Speech & Language*, vol. 27, no. 1, pp. 151–167, 2013.

[16] M. Wang, Y. Chen, Z. Tang, and E. Zhang, "I-vector based speaker gender recognition," in *2015 IEEE Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, 2015, pp. 729–732.

[17] M. Alhussein, Z. Ali, M. Imran, and W. Abdul, "Automatic gender detection based on characteristics of vocal folds for mobile healthcare system," *Mobile Information Systems*, vol. 2016, 2016.

[18] S.I. Levitan, T. Mishra, and S. Bangalore, "Automatic identification of gender from speech," in *Proceeding of Speech Prosody*, 2016.

[19] A. Larcher, K.A. Lee, B. Ma, and H. Li, "Text-dependent speaker verification: Classifiers, databases and RSR2015," *Speech Communication*, vol. 60, pp. 56 – 77, 2014.

[20] M. Sahidullah. and T. Kinnunen, "Local spectral variability features for speaker verification," *Digital Signal Processing*, vol. 50, pp. 1 – 11, 2016.

[21] M. Sahidullah and G. Saha, "Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition," *Speech Communication*, vol. 54, no. 4, pp. 543 – 565, 2012.

[22] N. Brümmer and E. de Villiers, "The bosaris toolkit: Theory, algorithms and code for surviving the new DCF," *arXiv preprint arXiv:1304.2865*, 2013.