DOI: 10.17469/O2108AISV000006

THAYABARAN KATHIRESAN

Gender bias in voice recognition: An i- and x-vector-based gender-specific automatic speaker recognition study

One of the critical implications of the physiological differences between adult males and females is acoustic differences in speech production. Such acoustic signal variability between the genders affects automatic speech processing applications, especially automatic speaker recognition systems. In this paper, the performance of the genders in state-of-the-art automatic speaker recognition algorithms, such as i- and x-vector, is studied by training the algorithms using a gender-balanced multilingual dataset and tested with gender-separated data from two different languages (English and Mandarin). Furthermore, generated i- and x-vector speaker embedding distributions in higher-dimensions are analysed using the t-SNE technique. The area distribution of speaker embeddings aids interpretation of the speaker recognition performances for both algorithms.

Keywords: speaker recognition, i-vectors, x-vectors, gender-difference, speaker-embeddings.

1. *Introduction*

Automatic speaker-recognition systems have emerged as an important means of verifying identity in many of today's e-commerce applications, as well as in general business interactions, forensics, and law enforcement (Hansen, Hasan, 2015; Kahn, Audibert, Rossato & Bonastre, 2010). An ideal speaker recognition system models the identity of a speaker and later verifies the claimed identity of said speaker using her/his spoken utterance in any adverse conditions. In practice, the complexity of the speaker recognition model and the performance of the system are influenced by speaker variability such as age, gender, language, health, etc. (González Hautamäki, Hautamäki & Kinnunen, 2019).

In a long stream of speaker recognition research, extrinsic and intrinsic variations in the speech signal are major challenges. The most common extrinsic variations include diversity in recording device, ambient acoustics, background noise, transmission channel, and distortions introduced in pre-processing algorithms (Nagrani, Chung, Xie & Zisserman, 2020).

For instance, initial research was constrained to text-dependent tasks and focused on solving the variation caused by pronunciation randomness, for which the Hidden Markov Model (HMM) was the most popular (Parthasarathy, Rosenberg, 1996). Later research attempted to solve text-independent tasks and had to deal with phonetic variation, which gave rise to the Gaussian Mixture Model with Universal Background Model (GMM-UBM) architecture (Reynolds, Quatieri & Dunn, 2000). Further research tried to address inter-session variation caused by

channels and speaking styles, for which the i-vector/PLDA architecture was the most successful (Dehak, Kenny, Dehak, Dumouchel & Ouellet, 2011). Recently, the research focus has been targeted towards dealing with complex variations in the wild scenarios, for which deep learning methods (deep neural network or DNN) have been demonstrated to be the most powerful (Okabe, Koshinaka & Shinoda, 2018; Snyder, Garcia-Romero, Sell, Povey & Khudanpur, 2018; Variani, Lei, Mcdermott, Moreno & Gonzalez-Dominguez, 2014).

Interestingly, however, performance degradation due to intrinsic variations, also known as within speaker variability, has received far less attention even though it has a strong impact on ASV system performance (González Hautamäki et al., 2019; Kahn et al., 2010). Within speaker variability arises from the speaker and can include changes in pronunciation, speaking style, short-term health condition, emotion, and/or vocal effort (Karlsson, Banziger, Dankovicová, Johnstone, Lindberg, Melin, Nolan & Scherer, 1998). Besides, biological differences between males and females have consequences for the sounds they produce, such as the inner dimension of the mouth, throat, and vocal folds (Simpson, 2009). It is also clear that we make specific speech patterns appropriate to the gender; for example, male vocal folds tend to be longer and thicker than female vocal folds causing them to vibrate more slowly. As a result, male speakers have an average F0 of 131 Hz (Hertz, cycles per second), and females produce approximately twice the male frequency (220 Hz) (Hillenbrand, Clark, 2009). In this paper, we limit our research focus to voice identity between genders.

To overcome the gender differences in recognition system performance, the speaker recognition community has concentrated on designing gender-conditioned systems. One such system was introduced in 2011 using a mixture of Probabilistic Linear Discriminant Analysis models (PLDA) with i-vector systems to make systems independent of speaker gender (Senoussaoui et al., 2011). The system was tested on 2010 NIST telephone speech (det5). The experiment showed that the Equal Error Rate (EER%) for male speakers was relatively better (1.81 EER%) than female speakers (2.47%). A pairwise discriminative training procedure for i-vector-based speaker recognition, presented in 2012, equally showed system performance for male speakers was relatively better in all the system variants (Cumani, Glembek, Brümmer, Villiers & Laface, 2012). Recent findings suggest that DNN-based speaker recognition methods such as x-vector systems achieve excellent results. The gender-dependent and independent systems' performance was tested using Kaldi-based x-vector techniques on SRE10 data, and the results show that the gender-dependent systems outperformed the independent systems (Snyder, Garcia-Romero & Povey, 2015). However, the gender-specific scores were not discussed in the paper.

In this current work, we study the impact of gender differences in i- and x-vector systems. In speaker recognition system design, we controlled the gender balance of data used in training the model and in testing the system. Moreover, we analyse speaker embeddings of both i- and x-vector systems using dimension

reduction techniques to understand the gender properties in high-dimensional embedding space.

The paper is organized as follows: Section 2 describes the i- and x-vectors, and we briefly recall the recognition scoring methods that have been used in the system evaluations. Section 3 presents the automatic speaker recognition experiment results and speaker embedding analysis. Finally, in Section 4, we discuss the findings and conclusion.

2. *Methods*

2.1 Dataset

We used the multilingual speaker recognition corpora Voxceleb1 (Nagrani, Chung & Zisserman, 2017) and Voxceleb2 (Chung, Nagrani & Zisserman, 2018) to train the speaker recognition models. In both the Voxceleb1 and Voxceleb2 datasets, the number of male speakers is higher than the number of female speakers. To balance the gender, we randomly selected male speakers to balance the number of female speakers. The original and modified speaker counts are shown in Table 1.

Table 1 - *Train dataset*

	Voxceleb1		Voxceleb2	
	Original	Modified	Original	Modified
#Total	1211	1092	5994	4402
#Male	665	546	3793	2201
#Female	546	546	2201	2201

For testing, we used two datasets, TIMIT (English) (Garofolo, Lamel, Fisher, Fiscus & Pallett, 1993) and AISHELL-1 (Mandarin) (Bu, Du, Na, Wu & Zheng, 2017). Similar to the modifications made to the training data, we did a gender balance in the test data, shown in Table 2.

Table 2 - *Test dataset*

	TIMIT		AISHELL-1	
	Original	Modified	Original	Modified
#Total	630	384	400	372
#Male	432	198	186	186
#Female	198	198	214	186

Though we used multilingual corpora for training the model, we used a language-specific (English and Mandarin) corpus for testing. The motivation is to explore the research question in a controlled speaker setting where we choose the language. In future studies, we will expand the question into more diverse and mixed language groups.

2.2 Automatic Speaker Recognition Algorithms

Both i- and x-vector systems were built using the Kaldi speech recognition toolkit (Povey, Ghoshal, Goel, Hannemann, Qian, Schwarz, Silovsk & Motl, 1968). We used a gender-balanced (see §2.1) version of Voxceleb1 and Voxceleb2 to train our recognition algorithms.

2.2.1 I-vector

An i-vector system[1] is a generative model that is derived using a total variability matrix (TVM) (Dehak et al., 2011). The TVM, obtained using unsupervised learning, is used to represent each utterance in a compact low-dimensional vector with an assumption that the speaker and session-dependent super vector M of Gaussian mean vectors may be modelled as

$$M = m + Tw \tag{1}$$

where m is the speaker and session-independent super vector obtained from a Gaussian mixture model (GMM) based universal background model (UBM), T is a low-rank total variability matrix that captures both speaker and session variability, and the i-vector is the posterior mean of w. The system is trained on 30 MFCC features with a frame-length of 25ms that are mean-normalized over a sliding window of up to 3 seconds. An energy-based speech activity detection (SAD) system selects features corresponding to speech frames. The UBM is a 2048 component full-covariance GMM. We extracted 400-dimensional i-vectors followed by an LDA scoring method (see §2.2.3).

2.2.2 X-vector

The x-vector system[2] is a time-delayed neural network architecture-based technique. We used 30-dimensional filter-banks with a frame-length of 25ms, mean-normalized over a sliding window of three seconds to train the x-vector system. Energy-based SAD (same as i-vector) is used to discard non-speech frames. We further applied augmentation on this data by adding music, speech, and noise using the MUSAN data set (Snyder, Chen & Povey, 2015); this consists of over 900 noises, 42 hours of music from various genres, and 60 hours of speech from twelve languages. This data augmentation helps the x-vector algorithm to be robust against noises and speech variabilities. Finally, we extracted 512-dimensional x-vectors followed by a PLDA scoring method (see §2.2.3).

2.2.3 Performance scoring

The results are presented in terms of equal error rate (EER), which corresponds to equal miss and false alarm rate. For both systems, linear discriminant analysis (LDA) is used to reduce the speaker embedding dimension. The LDA

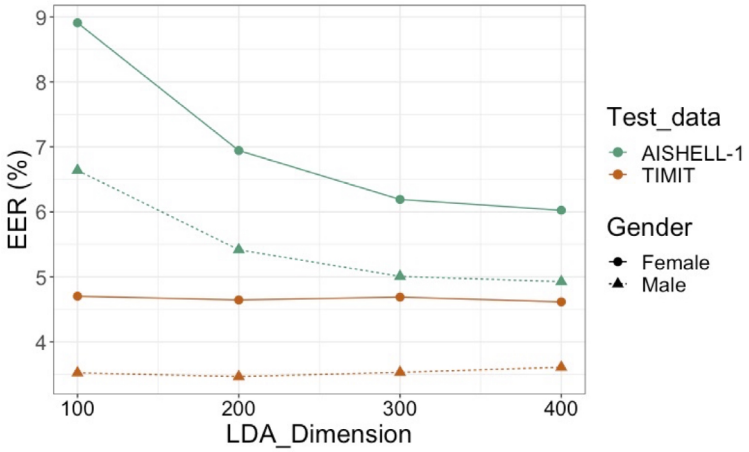
[1] Kaldi i-vector recipe at https://github.com/kaldi-asr/kaldi/tree/master/egs/voxceleb/v1

[2] Kaldi x-vector recipe at https://github.com/kaldi-asr/kaldi/tree/master/egs/voxceleb/v2

dimensionality varies from 100 to 400 for i-vectors and 100 to 500 for x-vectors with 100 step-size. Further, the probabilistic linear discriminant analysis (PLDA) model is used for channel/session compensation and measuring EER scoring in our experiments.

3. *Results*

Figure 1 - *The speaker recognition performance (EER%) of the i-vector system in 2 different datasets and different genders. The LDA -dimension is varied in the range of 100 to 400 with 100 step-sizes*



We trained the recognition models (i- and x-vectors) with undifferentiated gender (gender-balanced and mixed), and the testing was carried out on gender-separated datasets. Overall, the x-vector system outperformed the i-vector system, as expected. Notably, both the training methods are showing similar performance trends between genders. The male speaker recognition scores are better than those for female speakers in both i- and x-vector systems independent of dataset and LDA-dimensions. For i-and x-vectors systems, changes in LDA dimension improve the overall performance of the AISHELL-1 dataset but have no impact on TIMIT i-vector systems (Fig. 1), and show counter effects in TIMIT x-vector systems (Fig. 2). Using a t-distributed stochastic neighbour embedding (t-SNE) (Van der Maaten, Hinton, 2008) dimension reduction technique, the dimension of extracted speaker embeddings from both i-vectors (400-dimension) and x-vectors (512-dimension) were reduced to two-dimensional vectors, as shown in Figs. 3 and 4.

Figure 2 - *The speaker recognition performance (EER%) of the x-vector system in 2 different datasets and different genders. The LDA -dimension is varied in the range of 100 to 400 with 100 step-sizes*

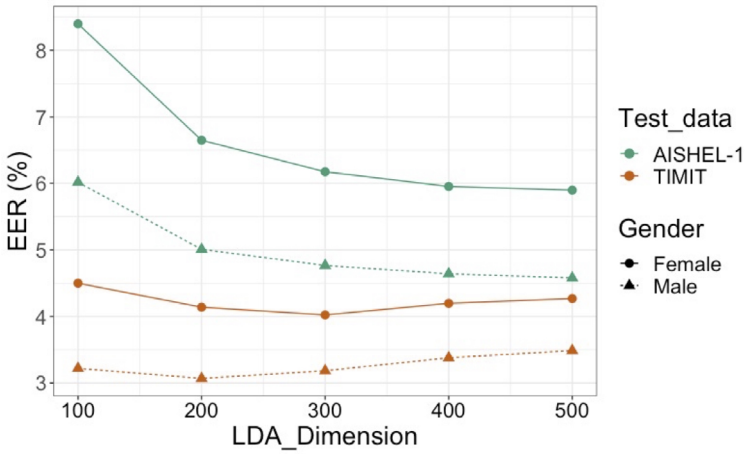
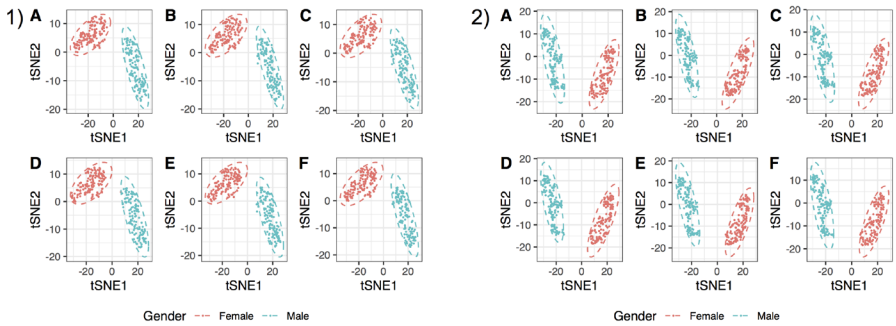


Figure 3 - *t-SNE reduced two-dimensional spaces of TIMIT speakers from i-vector (1) 400-dimension and x-vector (2) 512-dimension speaker embeddings*



The male and female speakers were clustered, and the cluster area is measured by fitting a 95% confidence ellipse. The elliptical area for male and female speakers in both TIMIT (Fig. 3) and AISHELL-1 (Fig. 4) datasets for i- and x-vectors are shown below. We did 6-fold cross-validation (sub-figs. A till F as shown in Figs 3 and 4) on each dataset and each recognition algorithm to get a statistically valid cluster area measure. The cluster area of 6-fold cross-validation measured in i- and x-vectors of TIMIT and AISHELL-1 dataset is shown in Fig. 5. A 3-way ANOVA (Area as dependent factor and Gender, Dataset, and Recognition Type as independent factors) was performed. The main effects of Dataset [$F(1,40) = 4.962$, $p=.0316$] and Type [$F(1,40) = 10.324$, $p=.0026$] are significant. The interaction is significant between Dataset and Type [$F(1,40) = 5.147$, $p=.0288$], and Gender and Type [$F(1,40) = 25.034$, $p<.0001$]. The three-way interaction is not significant [$F(1,40) = .796$, $p=.3775$].

Figure 4 - *t-SNE reduced two-dimensional spaces of AISHELL-1 speakers from i-vector (1) 400-dimension and x-vector (2) 512-dimension speaker embeddings*

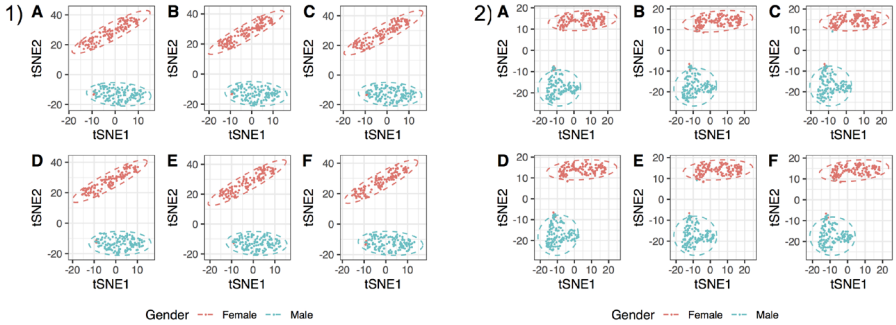
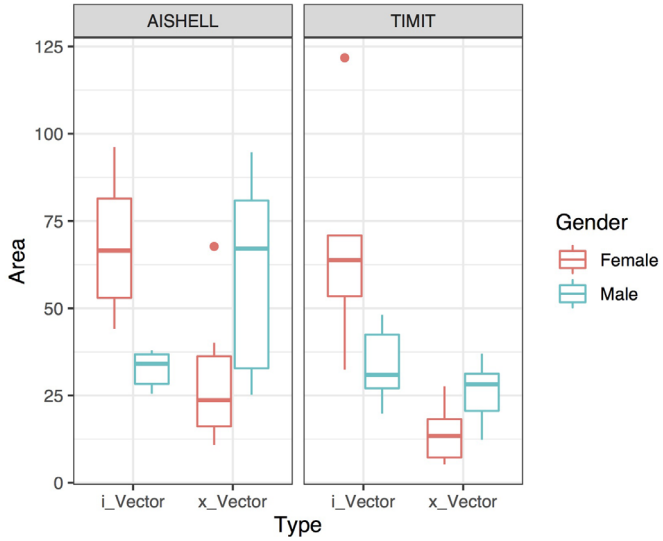


Figure 5 - *Cluster area of speaker embeddings from TIMIT and AISHELL-1 datasets in the 2-dimension t-SNE space*



4. *Discussion*

Overall, the findings show that the performance of male speaker verification is better than that for female speakers, independent of the voice recognition algorithm and language of the dataset. These gender performance differences therefore corroborate findings from previous studies on gender-dependent and independent speaker recognition results (Cumani et al., 2012; Senoussaoui, Kenny, Brümmer, De Villiers & Dumouchel, 2011; Snyder et al., 2015: 92-97). In the current investigation, we controlled parameters such as the gender balance in training and testing data and LDA dimensions; however, the recognition performance difference between the gender was unaffected. Furthermore, the performance of the x-vector system is better

than the i-vector system as expected (Snyder et al., 2018), since the power of DNN makes it possible to capture subtle speaker-specific indexical information.

The t-SNE based speaker embedding analysis sheds some light on understanding the distribution of gender in higher-dimensional spaces. The idea of finding a correlation between algorithm performance and speaker embedding distribution has shown different effects in two different algorithms. The male speaker embedding in the i-vector space occupies a smaller area than that occupied by the female speakers. In contrast, the female speakers in the x-vector embedding space occupy a smaller area than male speakers. Fundamentally, the idea and mathematics behind i-vector and x-vector algorithms are different. There is seemingly no reason that both algorithms would model the speakers similarly. However, they have shown comparable effects across two different datasets, which paves the way for understanding the speaker embedding distribution properties with speaker recognition performance.

In future research, the acoustic feature extraction parameters such as frame length and number of MFCCs used to build the recognition models will be investigated to further understand the performance difference between the genders. In addition, we will explore the implications of the area of embedding distributions in the higher dimension (see Fig. 5) by manipulating the spread. Alternatively, some new perspectives about speaker individualities in a voice, like how individual speakers control their identity to some degree, are recently being discussed (Dellwo, Pellegrino, He & Kathiresan, 2019). These will also be considered in further studies to investigate the gender difference in speaker recognition performance.

Acknowledgements

This work was supported by the Swiss National Science Foundation (SNSF), Grant No. 185399. I would like to thank Volker Dellwo for the support and supervision and Arjun Verma for his contributions in setting up the experiments. Also, I am grateful to Sandra Schwab for her contribution to the statistical analysis and Leah Bradshaw for proof reading.

Bibliography

Bu, H., Du, J., Na, X., Wu, B. & Zheng, H. (2017). AISHELL-1: An open-source Mandarin speech corpus and a speech recognition baseline. *20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment O-COCOSDA*, Seoul, South Korea, 1-3 September 2017, 1-5.

Chung, J.S., Nagrani, A. & Zimmerman, A. (2018). Voxceleb2: Deep speaker recognition. *Proceedings of the Annual Conference of the International Speech Communication Association*, Interspeech 2018, Hyderabad, India, 2-6 September 2018, 1086-1090.

Cumani, S., Glembek, O., Brümmer, N., De Villiers, E. & Laface, P. (2012). Gender independent discriminative speaker recognition in i-vector space. *IEEE International*

Conference on Acoustics, Speech and Signal Processing ICASSP, Kyoto, Japan, 25-30 March 2012, 4361-4364. doi: 10.1109/ICASSP.2012.6288885.

Dehak, N., Kenny, P.J., Dehak, R., Dumouchel, P. & Ouellet, P. (2011). Front-End Factor Analysis for Speaker Verification. In *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4), 788-798. https://doi.org/10.1109/TASL.2010.2064307.

Dellwo, V., Pellegrino, E., He, L. & Kathiresan, T. (2019). The dynamics of indexical information in speech: Can recognizability be controlled by the speaker?. In *AUC PHILOLOGICA*, 2, 57-75. https://doi.org/10.14712/24646830.2019.18

Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G. & Pallet, D.S. (1993). DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1. In *NASA STI/Recon Technical Report N*, 93.

Gonzaléz Hautamäki, R., Hautamäki, V. & Kinnunen, T. (2019). On the limits of automatic speaker verification: Explaining degraded recognizer scores through acoustic changes resulting from voice disguise. In *The Journal of the Acoustical Society of America*, 146(1), 693-704. https://doi.org/10.1121/1.5119240.

Hansen, J.H.L., Hasan, T. (2015). Speaker Recognition by Machines and Humans: A tutorial review. In *IEEE Signal Processing Magazine*, 32(6), 74-99. doi: 10.1109/MSP.2015.2462851

Hillenbrand, J M., Clark, M.J. (2009). The role of f0 and formant frequencies in distinguishing the voices of men and women. In *Attention, Perception, & Psychophysics*, 71(5), 1150-1166. https://doi.org/10.3758/APP.71.5.1150

Kahn, J., Audibert, N., Rossato, S. & Bonastre, J.F. (2010). Intra-speaker variability effects on Speaker Verification performance. *Odyssey – Speaker and Language Recognition Workshop*, Brno, Czech Republic, 28 June-1 July 2010, 109-116.

Karlosson, I., Banziger, T., Dankovicová, J., Johnstone, T., Lindberg, J., Melin, H., Nolan, F. & Scherer, K. (1998). Within-speaker variability due to speaking manners. In *5th International Conference on Spoken Language Processing ICSLP*, Sydney, Australia 30 November – 4 December 1998, 3(1), 2-5.

Nagrani, A., Chung, J.S., Xie, W. & Zisserman, A. (2020). Voxceleb: Large-scale speaker verification in the wild. In *Computer Speech and Language*, 60, 101027. doi: https://doi.org/10.1016/j.csl.2019.101027.

Nagrani, A., Chung, J.S. & Zisserman, A. (2017). VoxCeleb: A large-scale speaker identification dataset. *Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech 2017,* Stockholm, Sweden, 20-24 August 2017, 2616-2620.

Okabe, K., Koshinaka, T. & Shinoda, K. (2018). Attentive statistics pooling for deep speaker embedding. *Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech 2018*, Hyderabad, India, 2-6 September 2018, 2252-2256.

Parthasarathy, S., Rosenberg, A.E. (1996). General phrase speaker verification using sub-word background models and likelihood-ratio scoring. *Proceeding of Fourth International Conference on Spoken Language Processing ICSLP*, Philadelphia, USA, 3-6 October 1996, 2403-2406.

Povey, D., Ghosal, A., Goel, N., Hannemann, M., Qian, Y., Schwarz, P., Silovsk, J. & Motil, P. (1968). Sensory-perception testing box. In *Canadian Journal of Occupational Therapy*, 35(4), 140. PMID: 5195744.

Reynolds, D.A., Quatieri, T.F. & Dunn, R.B. (2000). Speaker verification using adapted Gaussian mixture models. In *Digital Signal Processing: A Review Journal*, 10(1), 19-41. https://doi.org/10.1006/dspr.1999.0361.

Senoussaoui, M., Kenny, P., Brümmer, N., De Villiers, E. & Dumouchel, P. (2011). Mixture of PLDA models in I-vector space for gender-independent speaker recognition. *Proceedings of the Annual Conference of the International Speech Communication Association*, *Interspeech 2011*, Florence, Italy, 27-31 August 2011, 25-28.

Simposon, A.P. (2009). Phonetic differences between male and female speech. In *Language and Linguistics Compass*, 3(2), 621-640. https://doi.org/10.1111/j.1749-818X.2009.00125.x

Snyder, D., Chen, G. & Povey, D. (2015). Musan: A music, speech, and noise corpus. In *ArXiv Preprint*: 1510.08484.

Snyder, D., Garcia-Romero, D. & Povey, D. (2015). Time delay deep neural network-based universal background models for speaker recognition. *IEEE Workshop on Automatic Speech Recognition and Understanding ASRU*, Scottsdale, Arizona, USA, 13-17 December 2015, 92-97.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D. & Khudanpur, S. (2018). X-Vectors: Robust DNN Embeddings for Speaker Recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP*, Calgary, Canada, 15–20 April 2018, 5329-5333.

Van Der Maaten, L., Hinton, G. (2008). Visualizing Data using t-SNE. In *Journal of Machine Learning Research*, 2579-2605.

Variani, E., Lei, X., Mcdermott, E., Moreno, I.L. & Gonzalez-Dominguez, J. (2014). Deep neural networks for small footprint text-dependent speaker verification. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP*, Florence, Italy, 4-9 May 2014, 4052-4056.