DATA AUGMENTATION ENHANCED SPEAKER ENROLLMENT FOR TEXT-DEPENDENT SPEAKER VERIFICATION

Achintya Kumar Sarkar1, Himangshu Sarma1, Priyanka Dwivedi1, Zheng-Hua Tan2

1 Indian Institute of Information Technology (IIIT), Sri City, Chittoor, India
2Department of Electronic Systems, Aalborg University, Aalborg, Denmark

sarkar.achintya@iiits.in

ABSTRACT

Data augmentation is commonly used for generating additional data from the available training data to achieve a robust estimation of the parameters of complex models like the one for speaker verification (SV), especially for under-resourced applications. SV involves training speaker-independent (SI) models and speaker-dependent models where speakers are represented by models derived from an SI model using the training data for the particular speaker during the enrollment phase. While data augmentation for training SI models is well studied, data augmentation for speaker enrollment is rarely explored. In this paper, we propose the use of data augmentation methods for generating extra data to empower speaker enrollment. Each data augmentation method generates a new data set. Two strategies of using the data sets are explored: the first one is to train separate systems and fuses them at the score level and the other is to conduct multi-conditional training. Furthermore, we study the effect of data augmentation under noisy conditions. Experiments are performed on RedDots challenge 2016 database, and the results validate the effectiveness of the proposed methods.

Index Terms: Data augmentation, Speaker enrollment, GMM-UBM, Noisy, Text-dependent Speaker verification

1. INTRODUCTION

Speaker verification (SV) [1] is defined as the task of verifying a person using their voice signal. It is a binary classification problem, where an SV system takes decision by either accepting or rejecting a person claiming the identity using his/her voice. As in most of machine learning methods, constructing an SV system consists of training and test phases. In the training/enrollment phase, speakers are characterized by their models/vectorized representation using his/her speech samples during training. In test, a speaker requests to grant the access of a system by claiming his/her identity with a voice sample. The delivered (test) speech sample is then scored against the claimant specific speaker representation in the system. Finally, the score is used for decision making whether the claimant will be accepted or rejected.

SV systems can be broadly divided into text-independent (TI) and text-dependent (TD). In TI-SV, speakers are free to speak any sentences during the enrollment and verification processes, whereas TD-SV constraints a speaker to speak a particular sentence during both the enrollment and verification/test phases. Since TD-SV maintains the matched phonetic contents between the enrollment and verification phases in contrast to the TI-SV, TD-SV yields lower error rates in speaker verification using short speech utterances. Therefore, TD-SV is suitable for real-time applications compared to TI-SV and is the focus of this paper.

There are many techniques available in the literature for the improvement of TD-SV systems using short utterances. Those techniques can be divided into different domains. For example, feature-domain approaches include Mel-frequency cepstral coefficients (MFCC) [2], perceptual linear prediction (PLP) [3], deep neural networks (DNNs) based Bottleneck feature [4], while model domain methods include Gaussian mixture models-universal background model (GMM-UBM) [5], i-vector [6] and x-vector [7] techniques.

In low resource applications, it is difficult to get a large amount of diverse data for training a large number of parameters in speaker independent model like GMM-UBM, DNNs, i-vector and x-vector. To create the diverse version of available training data, many augmentation techniques have been introduced in literature. Augmentation basically generates additional data from existing data with sort of transformations, for example, vocal tract length perturbation [8], mixing noise or other speech files with the given raw speech signal [9] [10], applying impulse (IR) response (of hall room, class room) on the given raw speech signal [11], quadrature distortion on raw audio signal [12], wow re-sampling [12], pitch shifting [13]. SpecAugment (deformation of log mel spectrogram with frequency masking) [14] and random image warping [15] on image. The effectiveness of data augmentation has been proven in various studies including speech recognition [14], speaker recognition [7] and image processing [16].

In speaker verification, augmented data, e.g., noisy version of available training data, are conventionally used to build speaker-independent (SI) modeling, e.g. GMM-UBM [17], DNNs [17], total variability space in i-vector [7], and in post-processing/scoring step e.g. probabilistic linear discriminate analysis (PLDA) [7] [19]. In [20] [17], the noisy version of training speech utterances/speaker enrollment data has been included in the enrollment phase for building a noise-robust model for spoofing detection [20] and speaker recognition [17] under noisy environments, respectively. However, as per our best knowledge, there is no study in the literature to use augmented data for speaker enrollment, other than creating a noisy version of speech data for the purpose of noise robustness. Therefore, it is interesting to investigate whether the class of data augmentation methods including pitch shifting, harmonic distortion, impulse response and mixing speech file are useful for speaker enrollment in TD-SV.

The main goal of this paper is to study the effect of different data augmentation techniques to increase the quantity of speaker enrollment data on the performance of speaker verification. We consider different strategies. First, speaker dependent models are trained for the particular augmentation method in the training phase and in the test phase, original evaluation data without augmentation are scored
Speaker enrollment phase:

Training data for speaker $r$

$\{1, \ldots, R\}$

Augmentation method −1

Feature extraction

Adaptation

Speaker model $\lambda_{\text{conv}}$

Augmentation method −2

Feature extraction

Adaptation

Speaker model $\lambda_{\text{aug}−1}$

Augmentation method −N

Feature extraction

Adaptation

Speaker model $\lambda_{\text{aug}−N}$

GMM−UBM

Test phase:

Test data with claimed as $r$

Feature extraction

Scoring

Average/Min/Max

Decision

Accept/Reject

all models $\epsilon r$

Speaker model $\lambda_{\text{con}}^r, \lambda_{\text{aug}−1}^r, \ldots, \lambda_{\text{aug}−N}^r$

Speaker model $\lambda_{\text{con}}^r, \lambda_{\text{aug}−1}^r, \ldots, \lambda_{\text{aug}−N}^r$

Fig. 1. Text-dependent speaker verification using augmentation and original training data

against the respective speaker models i.e. claimant specific models. It basically develops a separate SV system for each augmentation method. Score for a given test utterance from different systems are fused into a single value with average, maximum, minimum and median operations. Next, we also study the multi-conditioning training, where a speaker model is trained by pooling both augmentation data generated from different augmentation methods along with the original enrollment data in the database. The performance of TD-SV is studied on the RedDots challenge 2016 database [21] with the GMM-UBM technique. It is well know fact [22] that GMM-UBM yields lower error rate for speaker recognition using short utterances than the i-vector technique. Experimental results show that data augmentation reduces the error rates of TD-SV.

The paper is organized as follows: Section 2 describes the data augmentation methods. Section 3 describes the GMM-UBM technique for TD speaker verification. Experimental setup, and results and discussion are presented in Section 4 & Section 5 respectively. Finally, the paper is concluded in Section 6.

2. DATA AUGMENTATION METHODS

In this section, we briefly describe the different data augmentation techniques considered for TD-SV in this paper.

- Pitch shift [13]: In this method, the frequency of the voice/speech signal is either increased or decreased without affecting its duration. We consider two values, namely $\{1, 2\}$ in semitones, for pitch shifting of a given speech file in speaker’s enrollment data.

- Wow re-sampling [12]: This method is similar to the pitch shifting except for changing the intensity of the speech signal $x$ along the time,

$$\phi(x) = x + a \frac{\sin(2\pi f x)}{2\pi f}$$

where $\phi(x)$ is the transformed signal. $a$ and $f$ are the control parameters. We consider the change of minimum and maximum intensity and frequency values up to 3 and 2, respectively.

- Harmonic distortion [12]: It degrades the speech signal by applying $\sin(.)$ function on it multiple times. The value of degradation factor is considered to be 5 as per default parameters available in the toolbox.

- Impulse response (hall room) [11]: It modifies the given speech signal passing through a simulated source function/filter, e.g. acoustic impulse function/response of a class room. It can be thought as passing the speech signal though a filter which changes the input signal as per the characteristic of the responsive system.

- Sound mix [10]: It generates the modified speech signal by adding other audio files from within the same speaker. The generated speech will contain the attributes belonging to the same class. However, generated speech could be like as babble noise due to the overlapped of same person voice.

More details on augmentation techniques can be found in [12].

3. GMM-UBM TECHNIQUE

In this approach, a larger Gaussian mixture models (GMMs) [5] is trained using data from many non-target speakers called GMM-UBM. The GMM-UBM represents a large acoustic model space
which covers the various attributes available in the data. In the enrollment phase, speaker dependent models (of the registered speakers) are then derived from the GMM-UBM using the training/enrollment data for the particular speaker with maximum a posteriori (MAP) adaptation. In the test phase, the feature vectors of the test utterance $X = \{x_1, x_2, \ldots, x_T\}$ is aligned against the claimant $\lambda_r$ (obtained in enrollment phase) and GMM-UBM $\lambda_{ubm}$ models, respectively. Finally, a log-likelihood ratio value $\Lambda(X)$ is calculated using scores between the claimant and GMM-UBM models and is used to decide whether the claimant will be accepted or rejected.

$$\Lambda(X) = \frac{1}{T} \sum_{t=1}^{T} \left[ \log p(x_t|\lambda_r) - \log p(x_t|\lambda_{ubm}) \right]$$  \hspace{1cm} (2)$$

Fig.1 illustrates the TD speaker verification system using augmentation and original enrollment data (available in the database), where speaker models are trained using augmentation (generated from the speaker’s available enrollment data with augmentation methods) and original enrollment data, separately. Hence, it develops a number of TD-SV systems depending on the use of different enrollment data. In the test phase, scores of the test utterance from different TD-SV systems are fused with average/minimum/maximum operations shown in average/min/max block, respectively.

| Table 1. Number of trials available in RedDots evaluation condition for m-part01 task. |
| # of Genuine trials | # of trials in Non-target type |
|---------------------|-------------------------------|
|                      | Target -wrong | Imposter -correct | Imposter -wrong |
| 3242                | 29178          | 120086           | 1080774          |

4. EXPERIMENTAL SETUP

Experiments are performed on the male speakers’ parts (task m-part01) of the RedDots challenge 2016 database as per protocol [21]. There are 320 target (registered speakers) models to train and each has three sessions of recording speech samples. Utterances are very short in duration on an average of 2-3s duration. Database was collected in the different countries and then send to the other country through different networks to introduce the channel effect in speech signal. More details about the database can be found in [21]. Four types of test trials are available in the evaluation set to evaluate the system performance in Table1

- **Genuine trials**: when a target speaker speaks the passphrase/sentence in the test phase, which is same as used during the speaker enrollment phase
- **Target-wrong**: when a target speaker speaks a wrong (different) sentence in the testing phase as compared to their enrollment phase
- **Imposter-correct**: when an imposter speaks the same sentence as that of the target enrollment sessions
- **Imposter-wrong**: when an imposter speaks a wrong sentence in test phase as compared to the target enrollment pass-phrases

For signal processing, 57 dimensional MFCC [2](static $C_1$-$C_{19}$, $\Delta$, $\Delta\Delta$) using 25 ms of hamming window at frame rate of 10 ms. The MFCC features are then processed with RASTA filtering [23]. Afterward, robust voice activity detector (rVAD) [24] algorithm is applied to discard the less energized frames. Finally, selected frames are normalized to fit zero mean and unit variance normalization at utterance level. A gender independent GMM-UBM with 512 mixtures (having diagonal co-variance matrices) is trained using 6300 utterances from (438 males, 192 females) TIMIT database. 3 iterations are considered during MAP adaptation with the value of relevance factor 10. Audio degradation toolbox [12] is used to generate the augmentation data. To measure the performance of TD-SV, the equal error rate (EER) and minimum detection cost function (MinDCF) are used as per NIST 2008 SRE [25][26].

5. RESULT AND DISCUSSIONS

In this section, we analyze the performance of TD-SV system for different data augmentation methods (in the enrollment phase) and tested on original evaluation data with or without noise.

5.1. Effect on TD-SV performance using augmentation data for speaker enrollment when tested on clean evaluation data

Table 2 presents the comparison of TD-SV performance when speaker enrollment is done with or without data augmentation methods as well as various fusion strategies on the RedDot database (on task m-part01). Original indicates the speech files available in the database for training and testing. It is observed that the wow re-sampling augmentation method gives very close values of average EER/MinDCF compared to the baseline, i.e. the conventional system without data augmentation. This indicates that wow re-sampling generates content containing most speaker relevant information as compared with other augmentation techniques.

To further investigate the performance observed above, we plot the spectrograms of an original speech signal and the corresponding augmented data in Fig. 4. From Fig. 4 it is noticed that except for the wow re-sampling augmentation method, other methods significantly modify the structure of the spectrogram, especially on the higher frequency components (the most for IR method). These modifications are reflected in the performance of TD-SV. Now if we look at the score fusion among systems, fusion of $a$, $b$ and $f$ with the maximum method yields average EER of 2.46% and MinDCF of 1.04, both lower than those of the baseline system ($a$). This indicates the effectiveness of augmentation methods. Multi-condition $(a,b,f)$ presents the TD-SV with multi-condition training, where a speaker model is derived from GMM-UBM with MAP by pooling original enrollment data along with wow re-sampling, sound mix augmentation data. However, the error rate of the multi-condition training is slightly higher than the baseline. This indicates score fusion is a better choice for TD-SV using augmentation data.

5.2. Effect on TD-SV performance using augmentation data for speaker enrollment when tested on a noisy version of the evaluation data

Experimental results on the noisy version of the evaluation data are presented in Table 3. The noisy version of the evaluation data is generated as per ITU protocol [27]. For simplicity, we present only the system performance for market and car noise scenarios for SNR values of 5 and 10 dB, and maximum method (found optimal in previous subsection) in score fusion. The multi-condition method is not
Table 2. Comparison performance TD-SV for different enrollment data and fusion strategy on RedDots database (m-part01 task).

| System | Speaker Enrollment | Evaluation data | Non-target type [%EER/(MinDCF × 100)] | Average EER/MinDCF |
|--------|--------------------|-----------------|----------------------------------------|-------------------|
| a (Baseline) | Original | Original | 3.96/1.54 | 2.79/1.33 | 0.92/0.25 | 2.55/1.04 |
| b | Wow resampling | "" | 3.65/1.56 | 2.95/1.39 | 0.98/0.26 | 2.53/1.07 |
| c | Pitch shift | "" | 14.33/5.14 | 12.32/5.46 | 8.66/2.76 | 11.77/4.45 |
| d | Harmonic distort | "" | 14.68/4.97 | 11.48/4.33 | 8.36/2.56 | 11.51/3.95 |
| e | IR hall room | "" | 26.50/8.34 | 22.39/8.05 | 19.89/6.63 | 22.93/7.67 |
| f | Sound mix | "" | 8.07/3.01 | 6.45/2.87 | 3.96/1.17 | 6.16/2.35 |

| Score fusion | Method | Systems (a-f) | Average | Minimum | Maximum | Median |
|--------------|--------|---------------|---------|---------|---------|--------|
| Clean -original speech | Original | 4.34/1.68 | 3.08/1.44 | 1.60/0.43 | 3.01/1.18 |
| Wow resampling | Original | 4.53/1.88 | 3.39/1.69 | 1.41/0.42 | 3.11/1.33 |
| Harmonic distortion | Original | 16.03/5.96 | 11.25/5.18 | 9.37/3.46 | 12.22/4.87 |
| Sound mix | Original | 3.60/1.52 | 2.96/1.40 | 0.95/0.35 | 2.50/1.09 |
| Pitch shift | Original | 6.90/2.65 | 5.12/2.49 | 2.56/0.78 | 4.86/1.98 |
| IR hall room | Maximum | 3.67/1.53 | 2.83/1.33 | 0.89/0.24 | 2.46/1.04 |

Fig. 2. Spectrograms of the original and the corresponding augmented speech signals. The spoken content in the speech signal is "My voice is my password".

From Table 2 it can be seen that error rates of all systems significantly increase and are expected due to the mismatch between the enrollment with clean data and the evaluation with noisy data. Similarly to the Table 2, wow re-sampling and Sound mix show lower error rates than the other augmentation methods and fusion further improves the TD-SV with respect to the baseline. This indicates the usefulness of the data augmentation in TD-SV under noisy conditions.

6. CONCLUSION

In this paper, we proposed to explore a set of data augmentation approaches for generating extra data to empower speaker enrollment, in contrast to the use of data augmentation for building speaker independent models for TD-SV, in low resource applications. In the proposed method, each speaker is represented by a number of models that are derived from GMM-UBM using the original enrollment data together with augmented data (generated from the original enrollment data) for the particular speaker. It gives different TD-SV systems corresponding to different augmentation approaches. In the test, a test utterance is scored against different systems and then fused them with different strategies. Besides, we also evaluated the performance of TD-SV under clean and noisy environment conditions. Experimental results depicted that score fusion of the conventional/baseline system with the proposed data augmentation system reduces the error rate of TD-SV compared to their standalone counterpart. Experiments were conducted on the RedDots challenge 2016 database.

7. REFERENCES

[1] T. Kinnunen and H. Li, “An overview of text-independent speaker recognition: From features to supervectors,” Speech communication, vol. 52, no. 1, pp. 12–40, 2010.

[2] S. B. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continu-
Table 3. Comparison performance of TD-SV for different enrollment and evaluation data in RedDots database (m-part01 task).

| System | Speaker Enrollment | Evaluation Noise | SNR (dB) | Non-target type [%EER/(MinDCF × 100)] | Average EER/MinDCF |
|--------|--------------------|------------------|----------|--------------------------------------|-------------------|
|        |                    |                  |          | target-wrong | imposter-correct | imposter-wrong |              |
| a (Baseline) | Original | - | - | 3.96/1.54 | 2.79/1.33 | 0.92/0.25 | 2.55/1.04 |
|        | Market | 5 | 17.39/5.42 | 15.02/4.98 | 11.62/3.12 | 14.68/5.41 |
|        |        | 10 | 11.04/3.73 | 9.10/3.42  | 6.04/1.62  | 8.73/2.92  |
|        | Car     | 5 | 4.44/1.78  | 3.48/1.69  | 1.34/0.40  | 3.09/1.29  |
|        |        | 10 | 4.07/1.68  | 3.14/1.60  | 1.20/0.34  | 2.80/1.21  |
| b       | Wow resampling | Market | 5 | 16.64/5.25 | 15.02/5.10 | 11.50/3.18 | 14.38/4.51 |
|        |        | 10 | 10.79/3.67 | 9.38/3.50  | 5.89/1.66  | 8.69/2.95  |
|        | Car     | 5 | 4.41/1.70  | 3.79/1.79  | 1.57/0.42  | 3.25/1.30  |
|        |        | 10 | 3.85/1.59  | 3.39/1.68  | 1.30/0.35  | 2.85/1.21  |
| c       | Pitch shift | Market | 5 | 29.21/8.84 | 26.74/8.87 | 23.59/7.53 | 26.51/8.41 |
|        |        | 10 | 23.21/7.76 | 20.97/7.80 | 17.48/5.79 | 20.55/7.11 |
|        | Car     | 5 | 16.89/5.70 | 15.05/6.21 | 11.53/3.66 | 14.49/5.19 |
|        |        | 10 | 16.07/5.42 | 14.21/5.96 | 10.54/3.38 | 13.61/4.92 |
| d       | Harmonic distort | Market | 5 | 27.23/7.89 | 25.10/7.90 | 22.86/6.85 | 25.07/7.54 |
|        |        | 10 | 22.41/6.60 | 20.16/6.55 | 17.59/5.35 | 20.05/6.17 |
|        | Car     | 5 | 16.28/5.45 | 13.76/5.12 | 11.19/3.31 | 13.74/4.63 |
|        |        | 10 | 15.88/5.36 | 13.32/4.92 | 10.30/3.06 | 13.17/4.45 |
| e       | IR hall room | Market | 5 | 32.47/9.52 | 29.61/9.26 | 26.92/8.61 | 29.67/9.13 |
|        |        | 10 | 29.61/9.06 | 26.24/8.83 | 23.47/7.93 | 26.44/8.61 |
|        | Car     | 5 | 24.86/8.32 | 21.62/8.12 | 18.88/6.57 | 21.79/7.67 |
|        |        | 10 | 25.16/8.18 | 21.74/8.05 | 19.17/6.61 | 22.03/7.61 |
| f       | Sound mix | Market | 5 | 22.27/6.58 | 18.90/6.26 | 16.90/4.82 | 19.36/5.89 |
|        |        | 10 | 16.28/5.13 | 13.35/4.75 | 11.32/3.22 | 13.65/4.37 |
|        | Car     | 5 | 9.44/3.37  | 7.52/3.26  | 5.33/1.56  | 7.43/2.73  |
|        |        | 10 | 8.87/3.20  | 7.06/3.13  | 4.92/1.46  | 6.95/2.60  |
| Score fusion/(a,b,f) | Maximum | Market | 5 (dB) | 16.81/5.21 | 14.64/5.01 | 10.94/3.05 | 14.13/4.42 |
|        |        | 10 | 10.51/3.67 | 9.14/3.44  | 5.95/1.58  | 8.54/2.90  |
|        | Car     | 5 | 4.28/1.73  | 3.54/1.72  | 1.35/0.39  | 3.06/1.28  |
|        |        | 10 | 3.82/1.62  | 3.17/1.62  | 1.11/0.32  | 2.70/1.19  |

ousy spoken sentences,” *IEEE Trans. Acoust. Speech Signal Processing*, vol. 28, pp. 357–366, 1980.

[3] H. Hermansky, “Perceptual linear predictive (plp) analysis of speech,” *J. Acoust. Soc. Am.*, vol. 87, pp. 1738–1752, 1990.

[4] A. K. Sarkar, Z. H. Tan, H. Tang, S. Shon, and J. R. Glass, “Time-contrastive learning based deep bottleneck features for text-dependent speaker verification,” *IEEE/ACM Trans. Audio, Speech & Language Processing*, vol. 27, no. 8, pp. 1267–1279, 2019.

[5] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using adapted gaussian mixture models,” *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[6] N. Dehak, P. Kenny, R. Dehak, P. Ouellet, and P. Dumouchel, “Front-end factor analysis for speaker verification,” *IEEE Trans. on Audio, Speech and Language Processing*, vol. 19, pp. 788–798, 2011.

[7] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “X-vectors: Robust dnn embeddings for speaker recognition,” in *Proc. of IEEE Int. Conf. Acoust. Speech Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[8] N. Jaitly and G. E. Hinton, “Vocal tract length perturbation (vtlp) improves speech recognition,” in *In International Conference on Machine Learning (ICML)*, 2013.

[9] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Ng, “Deep speech: Scaling up end-to-end speech recognition,” in *arXiv*, 2014.

[10] M. Lasseck, “Audio-based bird species identification with deep convolutional neural networks,” in *In: Working Notes of CLEF 2018 (Cross Language Evaluation Forum)*.

[11] R. Stewart and M. Sandler, “Database of omnidirectional and b-format room impulse responses,” in *In Proc. IEEE Intl. Conf.*
[12] M. Mauch and S. Ewert, “The audio degradation toolbox and its application to robustness evaluation,” in International Society for Music Information Retrieval Conference, ISMIR, 2013, pp. 83–88.

[13] J. Salamon and J. P. Bello, “Deep convolutional neural networks and data augmentation for environmental sound classification,” IEEE Signal Processing Letters, vol. 24, no. 3, pp. 279–283, 2017.

[14] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, “SpecAugment: A simple data augmentation method for automatic speech recognition,” pp. 2613–2617, 2019.

[15] F. L. Bookstein, “Principal warps: Thin-plate splines and the decomposition of deformations,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 11, no. 6, pp. 567–585, 1989.

[16] C. Shorten and T. M. Khoshgoftaar, “A survey on image data augmentation for deep learning,” Journal of Big Data, 2019.

[17] R. Saeidi et al., “I4US submission to nist sre 2012: A large-scale collaborative effort for noise-robust speaker verification,” in Proc. of Interspeech, 2013, pp. 1986–1990.

[18] D. Michelsanti and Z.-H. Tan, “Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification,” in Proc. of Interspeech, 2017, pp. 2008–2012.

[19] A. K. Sarkar, D. Matrouf, P. M. Bousquet, and J. F. Bonastre, “Study of the effect of i-vector modeling on short and mismatch utterance duration for speaker verification,” in Proc. of Interspeech, 2012, pp. 2662–2665.

[20] H. Yu, A. Sarkar, D. A. L. Thomsen, Z. Tan, Z. Ma, and J. Guo, “Effect of multi-condition training and speech enhancement methods on spoofing detection,” in First International Workshop on Sensing, Processing and Learning for Intelligent Machines (SPLINE), 2016, pp. 1–5.

[21] “The reddots challenge: Towards characterizing speakers from short utterances,” https://sites.google.com/site/thereddotsproject/reddots-challenge.

[22] H. Delgado, M. Todisco, M. Sahidullah, A. K. Sarkar, N. Evans, T. Kinnunen, and Z.-H. Tan, “Further optimisations of constant q cepstral processing for integrated utterance and text-dependent speaker verification,” in Proc. of IEEE Spoken Language Technology Workshop (SLT), 2016.

[23] H. Hermanksy and N. Morgan, “Rasta processing of speech,” IEEE Trans. on Speech and Audio Processing, vol. 2, pp. 578–589, 1994.

[24] Z.-H. Tan, A. K. Sarkar, and N. Dehak, “rVAD: An unsupervised segment-based robust voice activity detection method,” Computer Speech & Language, vol. 59, pp. 1–21, 2020.

[25] A. Martin, G. Doddington, T. Kamm, M. Orodowski, and M. Przybocki, “The det curve in assessment of detection task performance,” in Proc. of Eur. Conf. Speech Commun. and Tech. (Eurospeech), 1997, pp. 1895–1898.

[26] “https://www.nist.gov/itl/iad/mig/2008-nist-speaker-recognition-evaluation-results,”.

[27] ITU, “G. 191: Software tools for speech and audio coding standardization,” 2005.