INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Variable frame rate-based data augmentation to handle speaking-style variability for automatic speaker verification

Amber Afshan[1], *Jinxi Guo*[1]*, *Soo Jin Park*[1]*, *Vijay Ravi*[1]*, *Alan McCree*[2], *and Abeer Alwan*[1]

[1]Dept. of Electrical and Computer Engineering, University of California, Los Angeles, USA
[2]Human Language Technology Center of Excellence, Johns Hopkins University, Baltimore, USA

amberafshan@g.ucla.edu

Abstract

The effects of speaking-style variability on automatic speaker verification were investigated using the UCLA Speaker Variability database which comprises multiple speaking styles per speaker. An x-vector/PLDA (probabilistic linear discriminant analysis) system was trained with the SRE and Switchboard databases with standard augmentation techniques and evaluated with utterances from the UCLA database. The equal error rate (EER) was low when enrollment and test utterances were of the same style (e.g., 0.98% and 0.57% for read and conversational speech, respectively), but it increased substantially when styles were mismatched between enrollment and test utterances. For instance, when enrolled with conversation utterances, the EER increased to 3.03%, 2.96% and 22.12% when tested on read, narrative, and pet-directed speech, respectively. To reduce the effect of style mismatch, we propose an entropy-based variable frame rate technique to artificially generate style-normalized representations for PLDA adaptation. The proposed system significantly improved performance. In the aforementioned conditions, the EERs improved to 2.69% (conversation – read), 2.27% (conversation – narrative), and 18.75% (pet-directed – read). Overall, the proposed technique performed comparably to multi-style PLDA adaptation without the need for training data in different speaking styles per speaker.

Index Terms: automatic speaker verification, speaking style, data augmentation, multicondition training

1. Introduction

An individual often varies his/her speaking style in day-to-day situations. Reading aloud, having a conversation, and talking to animals result in different acoustic properties in the speech signal. For instance, acoustical differences between read and conversational speech include different speaking rates and inconsistent pauses between words. There are also variations in the number and type of phonological phenomena observed. For example, vowels are modified or reduced in conversational speech, and word-final plosive bursts are not released while it is not the case in read speech [1]. Similar differences are observed across other speaking styles as well [2].

When the acoustic properties of an individual's speech differ between the enrollment and test utterances, automatic speaker verification (ASV) system performance generally degrades [3]. There are two categories of within-speaker variability that causes such difference: extrinsic and intrinsic variability. *Extrinsic variability* is associated with factors not directly related to the speaker's behavior (e.g., recording conditions, channel types, and environmental noise). There has been considerable progress in studying the effects of extrinsic variability on ASV performance [4, 5, 6, 7, 8, 9]. On the other hand, *intrinsic variability* is related to the speaker's conscious and/or unconscious behavior that can influence speech signal production. Studies showed that ASV performance degraded due to intrinsic variabilities–vocal effort, speaking styles, speaking rate, loudness, emotional state and physical status [10, 11, 12].

Speaking style variability is a type of intrinsic variability which can make acoustic characteristics considerably different within a speaker. However, only a limited number of studies have investigated the effects of style variability on ASV performance. *Style factors* are shown to be present in widely used speaker representations [13] such as i-vectors [14] and x-vectors [4]. ASV performance degradation due to style mismatch between the enrollment and test utterances were systematically analyzed in [15, 16, 17]. To alleviate the degradation due to style variabilities, some studies proposed the use of a joint factor analysis framework [11, 12]. In [18], curriculum-learning based transfer learning was done using neutral/physical stressed as well as read and spontaneous speech to compensate for style mismatches during testing. Note that the compensation techniques proposed in these studies require a variety of speaking styles per speaker to train the systems, i.e., the training data includes all the styles occurring in the test utterances [18]. However, one might not always have prior knowledge of the speaking style of the test utterances.

One can expect that including various speaking styles in the training data may improve the speaking-style robustness of the system. However, corpora with sufficient numbers of speakers speaking with different styles are not available. A widely-used approach to address insufficient data to train different conditions in ASV is *data augmentation* using artificially generated data. Augmentation strategies include adding variations of noise, reverberation [19, 4], collecting additional domain-specific data [18], and synthesizing data [20]. Yet, for style variability, artificially synthesizing speaking styles is not yet reliable enough to be applied [21, 13]. In this work, we propose the use of a variable frame rate (VFR) approach to generate style-normalized representations to perform data augmentation.

The rest of the paper is organized as follows. Section 2 describes the databases used. The proposed approach is detailed in Section 3. Section 4 provides the experimental setup and discusses the results, and we conclude with Section 5.

2. Databases

2.1. The UCLA Speaker Variability Database

In order to systematically study both within- and between-speaker variability, a multi-speaker speech database including multiple speech tasks per speaker is needed. The UCLA

[*] These authors have equal contribution

http://dx.doi.org/10.21437/Interspeech.2020-3006

Speaker Variability Database [22, 23] provides multiple recordings of speakers in a variety of speech tasks and on multiple occasions. Audio recordings were done in a sound-attenuated booth with a sampling rate of 22 kHz.

Speech tasks from the database used for this study include reading sentences to represent scripted speaking style ($\approx$ 75 sec); narrating a recent neutral, happy, or annoying conversation to represent unscripted affective speech ($\approx$ 30 sec each); making a telephone call to a familiar person to represent unscripted conversational style (60–120 sec); and talking aloud to pets in a video, providing pet-directed speech, which typically has exaggerated prosody (60–120 sec).

2.2. Databases for Training the ASV System

The Speaker Recognition Evaluation (SRE) databases developed by NIST are often used to train ASV systems. We used the NIST SRE 04, 05, 06, 08 and 10 databases [24, 25, 26] along with the Switchboard II corpus, phase 2 [27] for this purpose. The sampling rate for these databases is 8 kHz.

Note that although the SRE and Switchboard databases offer many recordings from a large number of speakers with multiple speech tasks, they do not provide multiple speech tasks per speaker under controlled recording environments. Additionally, they do not provide metadata regarding speaking style. Thus, the UCLA Speaker Variability Database is more suitable for detailed analyses of the effects of style variability. The UCLA dataset was downsampled to match the sampling rate of the training databases.

3. Proposed Approach

3.1. Automatic Speaker Verification System

The Kaldi [28] SRE16 recipe was used to develop a x-vector/PLDA ASV system [4]. The input acoustic features were 23-dimensional mel-frequency cepstral coefficients (MFCCs) with a frame length of 25 ms and a frame shift of 10 ms, which were mean normalized over a sliding window of up to 3 secs. Standard extrinsic data augmentation (as in the recipe) was applied on the training-data lists of both x-vector and PLDA.

A widely-used strategy to attenuate within-speaker variability is to train the PLDA with data for the conditions of variability from each speaker [29, 30]. Although this strategy has been mainly used for external sources of variability (e.g, noise, channel, etc.) [4, 30], it could be also applied to deal with the speaking style variability. However, sufficient amount of data is not available in the UCLA database to train a robust PLDA in this manner. Therefore, a PLDA model was trained with the previously mentioned training list and the in-domain adaptation (using the version provided in Kaldi) was performed with the UCLA database. The experimental configurations for adaptation will be described in Section 4.1.2.

3.2. Data Augmentation using Variable Frame Rate

In cases when multiple speaking styles per talker are not available in the training dataset, a method to artificially generate speaking style-normalized variants for augmentation is required. We propose to use the entropy-based variable frame rate to generate such variants. As mentioned earlier, some of the key differences across speaking styles are speech rate, long pauses, changes in the duration of individual sounds, boundary articulation. In this work, we aimed at reducing the effects of such acoustic differences on ASV performance. Specifically, we propose to generate style-normalized speaker representations by applying the entropy-based variable frame rate approach [31].

3.2.1. Entropy Computation

Consider a random variable $\nu \in \mathcal{R}^K$ where $p(\nu)$, the probability distribution function (PDF) of ν is a K-dimensional Gaussian. Let μ and Σ be the mean and covariance matrix of the random variable. The entropy can be calculated as:

$$\begin{aligned}
H(\nu) &= -\int p(\nu)\ln p(\nu)d\nu \\
&= -\int p(\nu)\left[-\frac{1}{2}(\nu-\mu)^T\Sigma^{-1}(\nu-\mu) - \ln|2\pi\Sigma|^{\frac{1}{2}}\right]d\nu \\
&= \frac{K}{2} + \frac{1}{2}\ln|2\pi\Sigma|
\end{aligned} \tag{1}$$

To facilitate faster computation and to avoid an ill-posed problem when the random variable's covariance matrix is not full rank [31], the following approximation is used to calculate the entropy:

$$H(\nu) \approx K\ln\sqrt{2\pi} + \ln \operatorname{Tr}\Sigma \tag{2}$$

3.2.2. Implementation

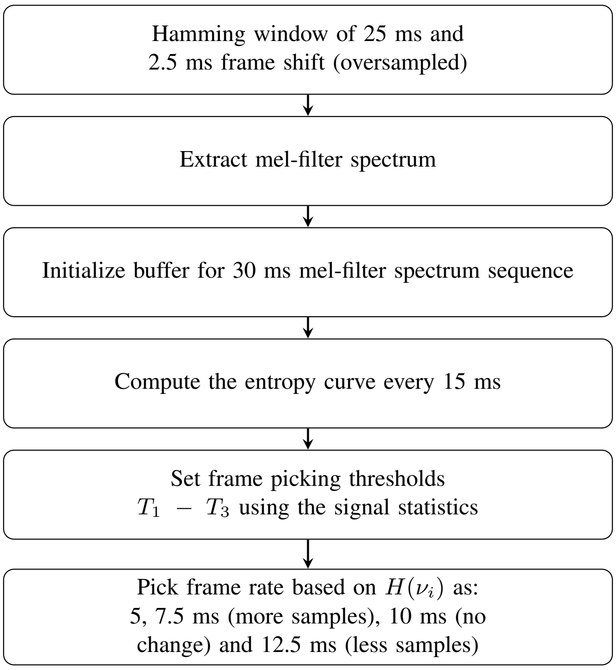
Hamming window of 25 ms and
2.5 ms frame shift (oversampled)
Extract mel-filter spectrum
Initialize buffer for 30 ms mel-filter spectrum sequence
Compute the entropy curve every 15 ms
Set frame picking thresholds
T_1 – T_3 using the signal statistics
Pick frame rate based on $H(\nu_i)$ as:
5, 7.5 ms (more samples), 10 ms (no change) and 12.5 ms (less samples)

Figure 1: *Overview of the entropy-based variable frame rate approach.*

The variable frame rate approach dynamically changes the frame rate based on inter-frame entropy using the steps shown in Figure 1. First, a signal is windowed using 25 ms Hamming window by first sampling with frame shift of 2.5 ms, a much lower value than widely-used 10 ms frame shift. With these densely sampled, or "oversampled" frames, varying frame rate becomes a simple task of retaining frames selectively. Mel-filter spectra are then computed. The frames spanning a duration of 30 ms are then used to calculate the entropy curve using the local entropy every 15 ms. VFR was carried out by comparing the signal's entropy to certain thresholds in order to calculate the frame picking rate in the extraction of MFCCs. Using the entropy curve of the speech signal $H(\nu_i)$, $i = 1, ..., N$, the frame-picking thresholds T_1, T_2, T_3 are set as in Equation 3.

$$\begin{cases} T_1 & = \omega_1 M_{max} + (1 - \omega_1) M_{med} \\ T_2 & = (1 - \omega_2) M_{max} + \omega_2 M_{med} \\ T_3 & = (1 - \omega_3) M_{med} + \omega_3 M_{min}, \end{cases} \quad (3)$$

where ω_1, ω_2, and ω_3 are weighting parameters of values 0.7, 0.8, and 0.5, respectively. M_{max}, M_{med}, and M_{min}, are the maximum, median, and minimum of the entropy curve, respectively. In this implementation, the x-vector extractor is trained using a frame shift of 10 ms. Hence, frame rates of 5 ms ($H(\nu_i) \geq T_1$) and 7.5 ms ($T_1 > H(\nu_i) \geq T_2$) are used to obtain more frames from the regions where the signal has rapid changes of information. A 10 ms frame shift is used when entropy is close to average ($T_2 > H(\nu_i) \geq T_3$). Whereas the frame rate is 12.5 ms ($T_3 > H(\nu_i)$) when the signal has low information gain, so that we obtain lesser frames from the region.

Fast speech rate and/or short pause can lead to a rapid change of information in spectral characteristics between frames resulting in a higher inter-frame entropy. Style variability may also cause a decrease in speech rate that could result in a low inter-frame entropy. That is, speaking style variability is at least partially reflected in the inter-frame entropy. Thus, extracting features maintaining consistent inter-frame entropy as much as possible could, in-turn, normalize the effects of style variability such as speech rate and duration of individual sounds.

Based on the above assumption, VFR was used to generate partially style-normalized utterance representations. This approach is expected to be more robust than varying the speaking rate of the entire utterance because the variations within an utterance and within speaking style are not always uniform due to speaker characteristics, context of the conversation, emotion, etc.

4. Experiments and Results

4.1. Experimental Setup

4.1.1. Database Statistics

A randomly selected subset of 50 female and 50 male speakers from the UCLA database was set aside as the "development set". The remaining subset of 50 female and 50 male speakers was used as the "evaluation set". The evaluation set was further split into "enrollment" and "test" set.

In order to analyze the effect of style variability on system performance, the effect of phonetic variability across utterances needs to be negligible. Based on the reports that 30-sec utterances cover enough phonetic variability to capture speaker-specific information [32], 30-sec long speech samples were used both for enrollment and test utterances. Table 1 shows the number of speech samples from the UCLA database used in this experiment. Note that at least 1 min of speech is required per speaker to generate style-matched enrollment – test utterance pairs. Because the majority of speakers in the UCLA database did not have enough speech in the narrative and pet-directed speaking styles, style-matched conditions for those styles were omitted. This resulted in 14 different evaluation combinations. All possible trials were generated for all the styles, which resulted in more non-target trials than target trials.

Table 1: *Number of utterances in distributed across each set for the UCLA database*

Speaking Style	**read**	**narrative**	**conversation**	**pet-directed**
Development set	196	36	184	19
Enrollment set	102	35	99	16
Test set	101	35	88	16

4.1.2. PLDA Adaptation Configurations

The PLDA trained on SRE and Switchboard data is adapted using the development set from the UCLA database. Recall that the major focus in this paper is data augmentation using VFR for PLDA adaptation. Hence, we designed the below five different adaptation configurations to experimentally analyze the advantages of the proposed technique:

Baseline: In-domain data with a single speaking style, the same as that of the enrollment set, is used (development set size X).

Extrinsic augmentation: Extrinsic variability is added using artificial data augmentation (development set size $5X$). The implementation here is similar to the one in x-vector training [4], but we use all the extrinsic variants and not a subset. We add music, noise and babble from the MUSAN corpus [33] and reverb by convolving with simulated room impulse responses [34].

VFR normalization: Entropy-based VFR normalization is applied to the development data of a single speaking style (development set size X). This generates partially style-normalized development set.

[Proposed] VFR normalization augmentation: Both the original representations of the development data and their style-normalized counterparts, obtained by performing VFR, were used (development set size $2X$).

Multi-style: Multiple speaking styles from the in-domain data were used (development set size $4X$).

In the baseline, extrinsic augmentation, VFR normalization, and VFR normalization augmentation configurations, the speaking style used in the development set matched that of the enrollment utterances. For instance, when enrolling with *read* and testing with other styles, the development set for PLDA adaptation contained only *read* sentences. In contrast, all styles in the development set were used in the multi-style configuration.

The baseline configuration was used to assess the effects of speaking style variability on ASV performance, as well as to establish baseline performance to be compared with the other configurations. The extrinsic augmentation configuration represents standard techniques that increase the amount of data, and it was used to understand how the proposed VFR data augmentation does when compared to it. The VFR normalization configuration was used to analyze the effectiveness of style-normalization with the VFR approach and also to assess if style-normalization alone would be enough to compensate for style variability.

Note that the multi-style configuration is the best-case scenario, but it is not realistic to assume that one can obtain all speaking styles for each speaker.

4.2. Results and Discussion

System performance in terms of the EER for the UCLA database is shown in Table 2. Statistical significance was verified using McNemar's test [35]. Unless mentioned explicitly, all performance differences reported in this section are significant with $p < 0.005$.

Table 2: *Performance in terms of EER (%) on the UCLA database.*

		Test			
	Enroll	**read**	**narrative**	**conversation**	**pet-directed**
Baseline	**read**	0.98	2.20	2.25	15.87
	narrative	0.63	NA	1.09	11.76
	conversation	3.03	2.96	0.57	22.12
	pet-directed	18.75	14.57	10.00	NA
Extrinsic aug.	**read**	0.98	1.89	3.37	12.50
	narrative	0.63	NA	1.09	11.76
	conversation	4.04	2.70	1.14	18.75
	pet-directed	12.50	13.73	10.00	NA
VFR norm.	**read**	0.98	1.89	3.37	18.75
	narrative	0.48	NA	1.09	11.76
	conversation	3.03	2.27	1.14	18.75
	pet-directed	12.50	15.69	13.33	NA
VFR norm. aug.	**read**	0.98	1.29	2.62	12.50
	narrative	0.63	NA	0.55	11.76
	conversation	2.69	2.27	0.38	18.75
	pet-directed	12.50	12.64	14.44	NA
Multi-style	**read**	0.98	1.26	2.25	12.50
	narrative	0.63	NA	0.73	11.76
	conversation	2.02	2.27	1.14	12.50
	pet-directed	12.50	15.59	13.33	NA

In the baseline, a style-mismatch between enrollment and test utterances consistently degraded ASV performance compared to their style-matched task. For instance, when enrolled with conversational speech, the style-matched task (conversation – conversation) had an EER of 0.57%. The performance degraded for style-mismatched tasks resulting in EERs of 3.03%, 2.96%, and 22.12% for conversation – read, conversation – narrative, and conversation – pet-directed pairs, respectively.

The second configuration of extrinsic augmentation performed better than the baseline in 6 tasks out of 14. These tasks were mainly the ones in which the development set was narrative or pet-directed speech. These styles had fewer utterances for adaptation and hence, the increase in the amount of data from augmentation could explain the improvement. On the other hand, the extrinsic augmentation performed worse than the baseline in 3 tasks out of 14 tasks. Interestingly, these were the tasks with reading or conversational speech as the development set. These styles had more utterances than others. The standard augmentation techniques used in the *extrinsic augmentation* setup merely increased the amount of data and might not have been sufficient to address style-variability.

VFR normalization was better than the baseline in 5 tasks out of 14, the same in 4 tasks out of 14, and worse in 5 tasks out of 14. This inconsistency in performance gains between the two setups may be due to: (i) the style normalization from VFR only partially addressed style variability (ii) the VFR normalization was only applied to development data and not to enrollment and test data. We did not apply VFR to enroll and test utterances because it would result in the loss of speaker-specific information.

The proposed approach of entropy-based VFR normalization augmentation performed better than the baseline in 9 tasks out of 14. The most notable improvement was seen when the testing was on pet-directed speech (read – pet-directed and conversation – pet-directed) which is often characterized by exaggerated prosody. However, for two tasks, read – conversation and pet-directed – conversation, the proposed approach did not improve the results compared to the baseline.

When compared to VFR normalization, the proposed approach showed significant improvement in 7 tasks out of 14. The performances were same in 5 tasks out of 14. There was a degradation in performance of the proposed approach for 2 tasks out of 14.

The proposed approach was better than extrinsic augmentation in 7 tasks out of 14 and the same in 6 tasks out of 14. The proposed approach was generally better even if it used less data than extrinsic augmentation. This result verifies the hypothesis that VFR, in fact, improved the ASV performance by providing partially style-normalized utterance representations and not by simply increasing the number of samples seen by the PLDA classifier. However, in the pet-directed – conversation task, characterized by exaggerated prosody, the proposed approach was worse than using extrinsic augmentation.

The multi-style configuration had more style information available in the development set as compared to the proposed approach–still, their performances were comparable. Their performances were the same in 6 tasks out of 14, 3 tasks out of 14 the proposed was better, and multi-style was better in 5 tasks out of 14. These findings support the hypothesis that VFR methods can be used as a data-augmentation technique when multi-style data are limited. One of the tasks where the proposed approach was better than multi-style was a style-matched task of conversation – conversation. There are probably variations within a speaking style that could be compensated by the style-normalized augmentation approach.

5. Conclusion

Speaking-style variability degraded ASV performance significantly. The proposed approach used an entropy-based variable frame rate technique to perform data-augmentation when multiple styles were not available to perform an in-domain adaptation of the PLDA classifier. The ASV performance showed significant improvement in the presence of a speaking-style mismatch by partially addressing performance degradation using VFR data augmentation. The performance of the proposed approach was comparable to the best-case scenario of having multiple styles available for PLDA augmentation. A natural progression of this work is to analyze other possible approaches to address the differences between speaking styles. It would also be interesting to investigate features and/or utterance representation techniques that are less affected by speaking style. More work will need to be done in the future to address the combined effects of speaking-style variability, short duration (< 30 secs), and extrinsic variability on ASV performance.

6. Acknowledgment

This work was supported in part by the NSF.

7. References

[1] M. A. Picheny, N. I. Durlach, and L. D. Braida, "Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech," *JSLHR*, vol. 29, no. 4, pp. 434–446, 1986.

[2] M. Eskenazi, "Trends in speaking styles research," in *Third European Conference on Speech Communication and Technology*, 1993.

[3] S. H. Shum, D. A. Reynolds, D. Garcia-Romero, and A. McCree, "Unsupervised clustering approaches for domain adaptation in speaker recognition systems," *Odyssey*, 2014.

[4] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *ICASSP*, 2018.

[5] J. Guo, R. Yang, H. Arsikere, and A. Alwan, "Robust speaker identification via fusion of subglottal resonances and cepstral features," *the Journal of the Acoustical Society of America*, vol. 141(4), pp. EL420–EL426, 2017.

[6] J. Guo, N. Xu, K. Qian, Y. Shi, K. Xu, Y. Wu, and A. Alwan, "Deep neural network based i-vector mapping for speaker verification using short utterances," *Speech Communication*, vol. 105, pp. 92–102, 2018.

[7] J. Guo, U. Nookala, and A. Alwan, "CNN-Based Joint Mapping of Short and Long Utterance i-Vectors for Speaker Verification Using Short Utterances," *Interspeech*, pp. 3712–3716, 2017.

[8] J. Guo, G. Yeung, D. Muralidharan, H. Arsikere, A. Afshan, and A. Alwan, "Speaker Verification Using Short Utterances with DNN-Based Estimation of Subglottal Acoustic Features," *Interspeech*, pp. 2219–2222, 2016.

[9] S. K. Sarangi and G. Saha, "Improved Speech-Signal Based Frequency Warping Scale for Cepstral Feature in Robust Speaker Verification System," *Journal of Signal Processing Systems*, Mar. 2020. [Online]. Available: https://doi.org/10.1007/s11265-020-01517-2

[10] E. Shriberg, M. Graciarena, H. Bratt, A. Kathol, S. S. Kajarekar, H. Jameel, C. Richey, and F. Goodman, "Effects of vocal effort and speaking style on text-independent speaker verification," in *9th Annual Conf. of the Intl. Speech Communication Association*, 2008.

[11] E. Shriberg, S. Kajarekar, and N. Scheffer, "Does session variability compensation in speaker recognition model intrinsic variation under mismatched conditions?" in *10th Annual Conf. of the Intl. Speech Comm. Assoc.*, 2009.

[12] S. Chen and M. Xu, "Compensation of Intrinsic Variability with Factor Analysis Modeling for Robust Speaker Verification," in *13th Annual Conf. of the Intl. Speech Communication Association*, 2012.

[13] J. Williams and S. King, "Disentangling Style Factors from Speaker Representations," *Interspeech*, pp. 3945–3949, 2019.

[14] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011, iSBN: 1558-7916.

[15] S. J. Park, C. Sigouin, J. Kreiman, P. A. Keating, J. Guo, G. Yeung, F.-Y. Kuo, and A. Alwan, "Speaker Identity and Voice Quality: Modeling Human Responses and Automatic Speaker Recognition," in *Interspeech*, 2016.

[16] S. J. Park, G. Yeung, J. Kreiman, P. A. Keating, and A. Alwan, "Using Voice Quality Features to Improve Short-Utterance, Text-Independent Speaker Verification Systems," *Interspeech*, 2017.

[17] S. J. Park, G. Yeung, N. Vesselinova, J. Kreiman, P. A. Keating, and A. Alwan, "Towards understanding speaker discrimination abilities in humans and machines for text-independent short utterances of different speech styles," *JASA*, vol. 144, no. 1, pp. 375–386, 2018.

[18] C. Zhang, S. Ranjan, and J. H. Hansen, "An Analysis of Transfer Learning for Domain Mismatched Text-independent Speaker Verification." in *Odyssey*, 2018, pp. 181–186.

[19] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," in *ICASSP*. IEEE, 2016, pp. 5115–5119.

[20] E. Rituerto-González, A. Mínguez-Sánchez, A. Gallardo-Antolín, and C. Peláez-Moreno, "Data Augmentation for Speaker Identification under Stress Conditions to Combat Gender-Based Violence," *Applied Sciences*, vol. 9, no. 11, p. 2298, 2019.

[21] Y. Wang, R. J. Skerry-Ryan, Y. Xiao, D. Stanton, J. Shor, E. Battenberg, R. Clark, and R. A. Saurous, "Uncovering Latent Style Factors for Expressive Speech Synthesis," *arXiv:1711.00520 [cs]*, Nov. 2017.

[22] J. Kreiman, S. J. Park, P. A. Keating, and A. Alwan, "The Relationship Between Acoustic and Perceived Intraspeaker Variability in Voice Quality," in *Interspeech*, Dresden, Germany, 2015.

[23] P. Keating, J. Kreiman, and A. Alwan, "A New Speech Database For Within- and Between-Speaker Variability," *Proc of the 19th ICPhS*, p. 4, 2019.

[24] M. Przybocki and A. Martin, "NIST Speaker Recognition Evaluation Chronicles," in *Proc. Odyssey*, 2004, pp. 12–22.

[25] M. Przybocki, A. Martin, and A. Le, "NIST Speaker Recognition Evaluation Chronicles - Part 2," in *Proc. Odyssey*, 2006, pp. 1–6.

[26] A. F. Martin and C. S. Greenberg, "NIST 2008 Speaker Recognition Evaluation: Performance across Telephone and Room Microphone Channels," in *Proc. Interspeech*, Brighton, UK, 2009, pp. 2579–2582, iSSN: 19909772.

[27] D. Graff, K. Walker, and A. Canavan, "Switchboard-2 phase ii," *LDC 99S79–http://www. ldc. upenn. edu/Catalog*, 1999.

[28] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, and others, "The Kaldi speech recognition toolkit," IEEE Signal Processing Society, Tech. Rep., 2011.

[29] D. Garcia-Romero, A. McCree, S. Shum, N. Brummer, and C. Vaquero, "UNSUPERVISED DOMAIN ADAPTATION FOR I-VECTOR SPEAKER RECOGNITION," *Odyssey*, p. 5, 2014.

[30] D. Garcia-Romero, X. Zhou, and C. Y. Espy-Wilson, "Multicondition training of Gaussian PLDA models in i-vector space for noise and reverberation robust speaker recognition," in *ICASSP*. IEEE, 2012, pp. 4257–4260.

[31] H. You, Q. Zhu, and A. Alwan, "Entropy-based variable frame rate analysis of speech signals and its application to ASR," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 2004, pp. I–549.

[32] T. Hasan, R. Saeidi, J. H. Hansen, and D. A. Van Leeuwen, "Duration mismatch compensation for i-vector based speaker recognition systems," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7663–7667.

[33] D. Snyder, G. Chen, and D. Povey, "MUSAN: A Music, Speech, and Noise Corpus," *arXiv:1510.08484 [cs]*, Oct. 2015.

[34] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *ICASSP*. IEEE, 2017, pp. 5220–5224.

[35] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947, publisher: Springer.