Learning from human perception to improve automatic speaker verification in style-mismatched conditions

Amber Afshan, Abeer Alwan

Department of Electrical and Computer Engineering, University of California Los Angeles, USA

amberafshan@g.ucla.edu, alwan@g.ucla.edu

Abstract

Our prior experiments show that humans and machines seem to employ different approaches to speaker discrimination, especially in the presence of speaking style variability. The experiments examined read versus conversational speech. Listeners focused on speaker-specific idiosyncrasies while "telling speakers together", and on relative distances in a shared acoustic space when "telling speakers apart". However, automatic speaker verification (ASV) systems use the same loss function irrespective of target or non-target trials. To improve ASV performance in the presence of style variability, insights learnt from human perception are used to design a new training loss function that we refer to as "C_{llr}CE loss". C_{llr}CE loss uses both speaker-specific idiosyncrasies and relative acoustic distances between speakers to train the ASV system. When using the UCLA speaker variability database, in the x-vector and conditioning setups, C_{llr}CE loss results in significant relative improvements in EER by 1-66%, and minDCF by 1-31% and 1-56%, respectively, when compared to the x-vector baseline. Using the SITW evaluation tasks, which involve different conversational speech tasks, the proposed loss combined with self-attention conditioning results in significant relative improvements in EER by 2-5% and minDCF by 6-12% over baseline. In the SITW case, performance improvements were consistent only with conditioning.

Index Terms: Style-robust, Speaker verification, Loss function, Conditioning, Attention

1. Introduction

Automatic speaker verification (ASV) is an open-set problem, i.e., test speakers are unavailable to the system during training but available during enrollment. ASV is, hence, a metric learning problem that maps speakers to a discriminative embedding space. Most of the work on speaker verification has focused on training with identification objectives. One such identification objective is cross-entropy loss [1, 2]. Identification loss functions learn linearly separable embeddings by focusing on maximizing inter-speaker distances. However, they do not typically minimize intra-speaker distances. Hence, the resulting embeddings do not have adequate discriminative properties.

To address the drawbacks of identification loss in ASV systems, Angular softmax [3] loss was used. Angular softmax uses cosine similarity as the logit input to the softmax layer. Additive margin variants of Angular softmax such as AM-Softmax [4, 5] and AAM-Softmax [6] use a cosine margin penalty on the target logit. These techniques although effective, have been proven sensitive to the value of scale and margin.

As an alternative to identification objectives, metric learning objectives that focus on minimizing intra-speaker distances have been used. Metric learning objectives such as contrastive loss [7] and triplet loss [8] have been used in ASV tasks with some success [9, 10]. However, these approaches require careful selection of triplet pairs i.e. anchor, positive and negative pairs, resulting in longer training cycles. Apart from the high computational cost, these losses do not consider the performance measures (such as equal error rate (EER) and detection cost function (DCF)) in training; these measures are used in the final evaluation of the speaker verification task.

It has been shown that considering a metric related to the final evaluation improves ASV performance further at least in text dependent ASV systems by using aAUC [11], aDCF [12] and C_{llr} [13] objectives. The C_{llr} loss, in particular, provides performance improvements without the need for triplet pairs and provides computational cost similar to that of identification objectives such as cross-entropy loss. C_{llr} was evaluated in a text dependent speaker verification task [13] and its efficacy has not been evaluated in a text independent case.

Given that everyday style variations in speech affect both inter- and intra-speaker variabilities [14, 7], it is important to use a loss function that maximizes inter-speaker distances and minimizes intra-speaker distances. To addresses this issue, in this paper, we introduce a loss function that is inspired by human speech perception.

1.1. Comparison between Humans and Machines

Speaking style variations occur frequently in everyday situations such as having a conversation, giving instructions, talking to a pet, etc. However, these variations have little effect on human ability to recognize a familiar voice [15]. Previously it has been shown that familiarity has an influence on human strategy to recognize talkers: familiar talkers are recognized by matching the stimuli to stored voice templates, while unfamiliar talkers are recognized through acoustic feature comparisons [16].

Humans have shown to outperform machines in a task of discriminating unfamiliar speakers in both style-matched and -mismatched conditions from samples of read and pet-directed speech (characterized by exaggerated prosody) [17, 18]. In our recent experiments [19], results suggest that humans and machines maybe employing different approaches to speaker discrimination in cases of moderate style variability. Moreover, two studies [20, 21] have shown that humans vary their perceptual strategies when "telling people together" versus "telling people apart." On the other hand, machines apply the same approach irrespective of target or non-target trials [18]. Given that humans and machines seem to employ different approaches to speaker discrimination, it is possible that machines might do better if they employed human perceptual strategies. In addition, humans might do better with machine assistance in certain situations. Therefore, we focus on learning from human speaker perceptual strategies in developing ASV algorithms, in particu-

This work was supported in part by the NSF.

lar, introducing a new training loss function.

In this work, we propose the C_{llr}CE loss function for text-independent ASV, especially in cases of style-mismatch. This loss function is inspired by strategies used by humans for an unfamiliar speaker discrimination task in the presence of moderate style variability (read versus conversational speech). Section 2 presents the proposed method. The experimental setup is described in Section 3, and the results and discussion are presented in Section 4. We conclude with Section 5.

2. Proposed Method

2.1. Human speaker perception

Our previous work [22] studied human speaker perception for moderate style variability (read versus conversational speech). The results showed that listeners find it easier to "tell speakers together" using speaker-specific idiosyncrasies, while listeners "tell speakers apart" based on relative positions within a shared acoustic structure rather than speaker-specific features.

This work aims to incorporate this strategy in the training loss function. Thus, we need a loss function that focuses on speaker-specific idiosyncrasies for the "target speaker" task while using acoustic distances between speakers for the "non-target speaker" task.

2.2. Embedding Extractors

An x-vector/PLDA system [23] is the baseline used in this paper. The inputs to the embedding extractor are 30-dimensional mel-frequency cepstral coefficients (MFCCs) using a 25 ms frame length and a 10 ms frame shift. The MFCCs are mean normalized over a sliding window of up to 3 secs. Extrinsic data augmentation of noise and reverberation [23] was applied to the training data.

Since, the x-vector system performance is degraded in the case of style-mismatch [24], we also want to evaluate the proposed method in a system that has lesser degradation due to style-mismatch. Hence, we perform additional experiments using an entropy-based variable frame rate (VFR) conditioning network [25, 26] developed to compensate for speaking style effects. This method uses VFR output [27, 24, 28] as a conditioning vector in the self-attention pooling layer. Five different approaches were used for conditioning. Among those, the best performing VFR conditioning network, concatenation with gating, is used. In this setup, the statistical pooling layer is replaced with a self-attention layer. The self-attention layer is then conditioned using an entropy-based variable frame rate vector [24].

2.3. Loss Functions

2.3.1. Cross-Entropy (CE) Loss

A widely-used loss function for training ASV systems, including the x-vector system, is the cross-entropy loss. This function calculates loss for a multi-class classification problem. CE loss can be calculated as,

$$L_{CE} = -\frac{1}{m}\sum_{i=0}^{m}\log\frac{e^{(\mathbf{W}_{\mathbf{y}_i}^T.\mathbf{x}_i+\mathbf{b}_{\mathbf{y}_i})}}{\sum_{j=0}^{N}e^{(\mathbf{W}_j^T.\mathbf{x}_j+\mathbf{b}_j)}} \quad (1)$$

where $\mathbf{x}_i$ is the i^{th} training sample, $\mathbf{y}_i$ is the ground truth speaker label of the i^{th} training sample, $i \in \{1, \ldots, m\}$, where m is the total number of training samples. $\mathbf{W}$ indicates the weight matrix, $\mathbf{b}$ is the bias vector. $\mathbf{W}_j$ and $\mathbf{W}_{\mathbf{y}_i}$ are the j^{th} and $\mathbf{y}_i^{\text{th}}$ columns of $\mathbf{W}$, respectively. $\mathbf{b}_j$ and $\mathbf{b}_{\mathbf{y}_i}$ are the j^{th} and $\mathbf{y}_i^{\text{th}}$ bias values, respectively. The CE loss is calculated for a total of N speakers.

The CE loss aims at maximizing inter-speaker distances but it does not minimize intra-speaker distances. By maximizing inter-speaker distances (the posterior probability of the correct class), the extracted embeddings are linearly separable. On the other hand, for the embeddings to include desirable discriminative features, the loss should also minimize intra-speaker distances (that is increase embedding similarity). The embeddings trained on CE loss–maximizing inter-speaker distances–are equivalent to the human approach of focusing on relative positions within a shared acoustic structure to "tell speakers apart". To minimize intra-speaker distances and implement other aspects of human perception strategies, we need a loss that focuses on speaker-specific idiosyncrasies.

2.3.2. C_{llr} Loss

To focus on speaker-specific idiosyncrasies without increasing the length of the training cycles, we chose the log-likelihood-ratio cost function (C_{llr}) [29] as a loss function for training the embedding extractor that we refer to as "C_{llr} loss".

C_{llr} is an application independent measure for evaluating soft decisions in ASV performance. There is a closed-form solution for C_{llr} [29] that provides the C_{llr} loss function as follows:

$$C_{\text{llr}}(\theta) = \frac{1}{2}\left(\frac{C_{tar}(\theta)}{N_{\text{tar}}} + \frac{C_{non}(\theta)}{N_{\text{non}}}\right) \quad (2)$$

$$C_{tar}(\theta) = \sum_{i\in\text{tar}}\log_2(1+e^{-s_\theta(\mathbf{x}_i,\mathbf{y}_i)}) \quad (3)$$

$$C_{non}(\theta) = \sum_{i\in\text{non}}\log_2(1+e^{s_\theta(\mathbf{x}_i,\mathbf{y}_i)}) \quad (4)$$

where θ represents the model parameters, $s_\theta(\mathbf{x}_i, \mathbf{y}_i)$ is the score from the last layer of the embedding extractor for speaker $\mathbf{y}_i$ from input $\mathbf{x}_i$, 'tar' is a set of target speakers and 'non' is a set of non-target speakers. The two terms in Equation 2 represent the costs for N_{tar} "target" ($C_{tar}(\theta)$) and N_{non} "non-target" speakers ($C_{non}(\theta)$).

C_{llr} can be interpreted as a measure that is inversely related to information. The lower the C_{llr}, the more the average information per trial (in bits) increases. Optimization is performed with the objective of minimizing C_{llr} loss. C_{llr} loss is calculated for each minibatch by considering the outputs of the last linear layer as scores and using the class labels to define target and non-target speakers. Thus, C_{llr} loss minimizes intra-speaker distances by focusing on speaker-specific idiosyncrasies. This is similar to the human approach to "tell speakers together".

2.3.3. Proposed method: C_{llr}CE loss

We propose to use the combination of cross-entropy loss and C_{llr} loss for training ASV systems, so that the loss function can maximize "inter-speaker" distances and minimize "intra-

Table 1: *UCLA SVD database statistics in terms of number of utterances.*

Style	read	instructions	narrative	conversation	pet-directed
Enroll	200	204	625^+	197	35^+
Test	199	204	625^+	174	35^+

$^+$ Same enroll and test utterances.

Table 2: *Performance using the UCLA database (in EER and minDCF) with CE and C_{llr}CE loss functions. The loss functions are used to train the x-vector system and the best performing VFR conditioning: concatenation with gating. The best performance in each condition with a statistically significant improvement over the baseline is boldfaced. If denoted by a '*' it is not a statistically significant improvement over the baseline.*

Loss		**CE**				**C_{llr}CE**			
Enroll	**Test**	**x-vector (Baseline)**		**VFR conditioning**		**x-vector**		**VFR conditioning**	
		EER %	**minDCF$_{0.01}$**	**EER %**	**minDCF$_{0.01}$**	**EER %**	**minDCF$_{0.01}$**	**EER %**	**minDCF$_{0.01}$**
read	read	0.50	0.018	0.50	0.013*	0.50	0.023	0.50	0.018
	instructions	0.49	0.054	0.49	0.037*	0.49	0.037*	0.49	0.027*
	conversation	2.86	0.254	2.29	0.232	2.29	0.240	**1.71**	**0.197**
	narrative	0.80	0.162	0.80	0.115*	0.80	0.123*	0.80	0.104*
	pet-directed	17.14	0.928	**14.29**	0.943	17.14	0.943	17.14	0.886*
instructions	read	1.47	0.154	**0.98**	0.120	1.47	0.137	1.47	0.108*
	instructions	0.45	0.005	0.45	0.005	0.45	0.005	0.45	0.005
	conversation	2.79	0.296	2.79	0.263*	2.79	0.263*	**2.24**	**0.238**
	narrative	1.23	0.110	**0.77**	0.102	0.92	0.090	**0.77**	**0.072**
	pet-directed	18.92	0.933	**13.51**	0.933	16.22	0.920	16.22	**0.908**
conversation	read	2.03	0.246	**1.52**	0.178	**1.52**	0.188	**1.52**	**0.173**
	instructions	2.97	0.267	2.48*	0.248*	2.48*	0.225*	2.48*	0.213*
	conversation	0.57	0.035	0.57	0.035	0.57	0.029*	0.57	0.020*
	narrative	1.94	0.224	1.94	0.187	1.94	0.179	2.10	**0.155**
	pet-directed	20.00	0.887	**17.14**	0.915	**17.14**	0.900	**17.14**	**0.858**
narrative	read	0.48	0.046	0.32*	0.032*	**0.16**	0.036	**0.16**	**0.020**
	instructions	0.46	0.024	0.46	0.019*	0.46	0.019*	0.46	0.013*
	conversation	1.46	0.132	1.10	0.121	1.10	0.127	**0.73**	**0.096**
	pet-directed	18.58	0.828	**13.27**	0.908	13.27	0.855	14.16	**0.841**
pet-directed	read	14.29	0.886	14.29	0.829*	14.29	0.857*	14.29	0.871*
	instructions	18.92	0.919	**13.51**	0.946	16.22	0.934	16.22	**0.908**
	conversation	21.21	0.914	**18.18**	0.867	**18.18**	0.842	21.21	0.774*
	narrative	19.47	0.886	**14.16**	0.929	15.93	0.892	15.04	**0.864**

speaker" distances. We thus use a combined loss function and refer to it as "C_{llr}CE loss",

$$C_{\text{llr}}CE(\theta) = \frac{1}{2}\left(C_{\text{llr}}(\theta) + L_{CE}\right) \tag{5}$$

Given that this loss function is inspired by human speaker discrimination strategies in the presence of moderate style variability, i.e, between read and conversational speech, we hypothesize that this loss function will provide the most improvement in conversational speech tasks.

3. Experimental Setup

Experiments were setup using Pytorch [30] and Kaldi [31]. Adam [32] optimization was used with a batch size of 128 and trained for 100 epochs.

3.1. Databases

3.1.1. The UCLA Speaker Variability Database (SVD)

To systematically study performance in the presence of style variability, the UCLA Speaker Variability Database [33, 34, 35, 36], a multi-speaker speech database including multiple speech tasks per speaker is employed. It incorporates commonly-occurring variations in speech from 101 female and 101 male speakers, recorded in a sound-attenuated booth at a sampling rate of 22kHz. The tasks include **reading** sentences characterizing scripted speaking style ($\approx$ 75 sec per speaker); giving **instructions** as unscripted clear monologue style ($\approx$ 30 sec per speaker); **narrating** a recent happy, annoying, or neutral conversation characterizing unscripted affective speech ($\approx$ 30 sec each affect per speaker); having a conversation on a call with a familiar person (speaker's side speech only) characterizing unscripted **conversational** style (60–120 sec per speaker); and talking to pets in a video representing **pet-directed** speech (60–120 sec per speaker).

To cover enough phonetic variability, such that there is negligible effects from it, and style variability is predominant [37], 30 sec-long speech samples were used for evaluation, This results in a total of 1,838 30 sec segments for evaluation as shown in Table 1. A majority of speakers had less than 1 min of speech for pet-directed speech and affect-matched narrative case. Hence, the style-matched cases for those styles were omitted, as a style-matched case requires at least 1 min (two 30 sec samples) of speech from the same speaker. This provides a total of 23 style-matched and mismatched tasks for evaluation. The UCLA SVD data were downsampled to 16 kHz, to match the rest of the databases used.

3.1.2. The Speakers in the Wild Database (SITW)

To evaluate the performance of the proposed loss on a large-scale database we use SITW [38] for evaluation. It includes speakers employing multiple speaking styles such as interviews, presentation, talk show, social-media videos etc. This database consists of 2,883 recordings from 117 male and 63 female speakers divided into 6,445 utterances sampled at 16 kHz. Single-speaker utterances in the eval set are referred to as "core". Enrollment utterances with multiple speakers (segmentation labels for the person of interest (POI) available) are re-

Table 3: *Performance using the SITW evaluation (in EER and minDCF) with CE, C_{llr}, and C_{llr}CE loss functions. The loss functions are used to train the x-vector system and the best performing VFR conditioning: concatenation with gating. The best performance in each condition with a statistically significant improvement over the baseline is boldfaced.*

		Core-Core		Core-Multi		Assist-Core		Assist-Multi	
Loss	**Model**	**EER %**	**minDCF$_{0.01}$**	**EER %**	**minDCF$_{0.01}$**	**EER %**	**minDCF$_{0.01}$**	**EER %**	**minDCF$_{0.01}$**
CE	x-vector (Baseline)	3.66	0.3820	5.87	0.4629	5.47	0.4041	6.90	0.4512
	VFR conditioning	3.69	0.3989	5.81	0.4740	**5.26**	0.4027	**6.54**	0.4651
C_{llr}	x-vector	4.13	0.4153	6.46	0.4940	6.24	0.4376	7.57	0.4824
	VFR conditioning	4.29	0.4009	6.65	0.4821	6.28	0.4337	7.68	0.4776
C_{llr}CE	x-vector	3.77	0.3654	5.88	0.4394	5.70	0.3833	6.74	0.4290
	VFR conditioning	**3.47**	**0.3346**	**5.73**	**0.4178**	5.36	**0.3738**	6.73	**0.4191**

ferred to as "assist", while the test utterances that do not include segmentation labels for POI are referred to as "multi".

3.1.3. VoxCeleb Database

ASV systems were trained on the Voxceleb2 *DEV* set [39]. It consists of speech from YouTube videos of 3,682 male and 2,313 female speakers and includes 1,092,009 utterances with a sampling rate of 16 kHz. The main disadvantage of using VoxCeleb2 for testing is that it comprises interview-style speech only and does not include different styles for each speaker. Hence, we believe that this database does not provide a good representation of the test case scenario targeted in this work.

4. Results and Discussion

4.1. UCLA SVD Evaluation

The loss functions used in our experiments are cross-entropy loss (CE), C_{llr} loss, and C_{llr}CE loss. These loss functions are used to train the x-vector system and the best performing VFR conditioning: concatenation with gating. Table 2 compares the performance (in EER % and minDCF) of the CE and C_{llr}CE loss functions for the UCLA database. The C_{llr} loss function by itself does not provide an improvement over the widely-used CE loss function in both the x-vector and VFR conditioning architectures. Therefore, we do not report those results in Table 2. Statistical significance was verified using McNemar's test [40]. Unless mentioned explicitly, all performance differences reported in this section are significant with $p < 0.05$.

In the x-vector setup, C_{llr}CE loss provides statistically significant improvements over CE loss in 11/23 tasks and is the same as CE loss in 12/23 tasks. The minDCF is significantly better with C_{llr}CE loss in 7/23 tasks and worse in 4/23 tasks when compared to CE loss. In the VFR conditioning setup, C_{llr}CE loss provides statistically significant improvements in 4/23 tasks, especially in tasks involving conversational style, compared to VFR conditioning with CE loss. Since in the VFR conditioning setup style variability is addressed, the improvement with C_{llr}CE loss is not consistent. The CE loss performs significantly better in 8/23 tasks. The performance in terms of minDCF values show that the C_{llr}CE loss provides significant improvements over CE loss in 12/23 tasks, and the same performance in 11/23 tasks. The most relative improvement is seen with tasks that include conversational or narrative style speech in enrollment and/or test conditions.

Overall, VFR conditioning trained with C_{llr}CE loss provides significant improvements over the x-vector baseline (with CE loss) in 7/23 tasks in EER, and 12/23 tasks in minDCF. When compared to their CE counterparts, we again notice that the conditions where the C_{llr}CE loss provides improvements are the ones that include conversation style speech and narrative style speech (closest to the conversation style).

4.2. SITW Evaluation

Table 3 presents the performance on the SITW evaluation set using different loss functions. The loss functions used are cross-entropy loss (CE), C_{llr} loss, and C_{llr}CE loss. These loss functions are used to train the x-vector system and the best performing VFR conditioning: concatenation with gating.

The results show that the best performing system in terms of minDCF values is the one with combination loss in the VFR conditioning setup. However, EER values of the C_{llr}CE loss in the VFR conditioning setup are slightly worse than the CE loss counterpart for assist-core and assist-multi evaluations. Overall, the proposed loss function with the VFR conditioning setup results in the best performance on the SITW evaluation. Since SITW involves mainly conversational speech, this result agrees with our hypothesis that the new loss function improves ASV system performance for conversational styles.

Overall results show that the combined C_{llr}CE loss improves ASV performance for the two configurations when compared to CE and C_{llr} loss functions individually. Thus, implying that the C_{llr} and CE loss functions are complementary.

5. Conclusion

In order to improve ASV performance in the presence of style variability, this work introduces a new loss function (C_{llr}CE) that is inspired by human speech perception. C_{llr}CE loss focuses on both speaker-specific idiosyncrasies to "tell speakers together" and on relative acoustic distances between the speakers to "tell speakers apart". This combined loss maximizes inter-speaker distances while minimizing intra-speaker distances resulting in performance improvements over the widely used CE loss function and also the C_{llr} loss function, showing their complementarity. To the best of our knowledge, this is the first work to propose a training loss function for ASV that is inspired by human perception. In future, this work will be extended to study perception strategies between other styles and use those to improve ASV approaches. Further studies on the effects on short-duration scenario [41, 42] and other embedding extractors [43, 44] would provide better understanding of the proposed loss function.

6. References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Adv Neural Inf Process Syst*, vol. 25, 2012.

[2] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[3] Y. Li, F. Gao, Z. Ou, and J. Sun, "Angular Softmax Loss for End-to-end Speaker Verification," in *ISCSLP*, 2018, pp. 190–194.

[4] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "CosFace: Large Margin Cosine Loss for Deep Face Recognition," in *CVF*, 2018, pp. 5265–5274.

[5] F. Wang, W. Liu, H. Liu, and J. Cheng, "Additive Margin Softmax for Face Verification," *IEEE SP Letters*, vol. 25, no. 7, pp. 926–930, 2018, arXiv: 1801.05599.

[6] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *CVPR*, 2019.

[7] S. Chen and M. Xu, "Compensation of Intrinsic Variability with Factor Analysis Modeling for Robust Speaker Verification," in *Interspeech*, 2012.

[8] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *CVPR*, Boston, MA, USA, 2015, pp. 815–823.

[9] C. Zhang, K. Koishida, and J. H. L. Hansen, "Text-Independent Speaker Verification Based on Triplet Convolutional Neural Network Embeddings," *IEEE-ACM T AUDIO SPE*, vol. 26, no. 9, pp. 1633–1644, Sep. 2018.

[10] F. A. R. R. Chowdhury, Q. Wang, I. L. Moreno, and L. Wan, "Attention-Based Models for Text-Dependent Speaker Verification," *arXiv:1710.10470 [cs, eess, stat]*, Jan. 2018.

[11] V. Mingote, A. Miguel, A. Ortega, and E. Lleida, "Optimization of the area under the ROC curve using neural network supervectors for text-dependent speaker verification," *Comput Speech Lang*, vol. 63, p. 101078, Sep. 2020.

[12] V. Mingote, A. Miguel, D. Ribas, A. Ortega, and E. Lleida, "Optimization of False Acceptance/Rejection Rates and Decision Threshold for End-to-End Text-Dependent Speaker Verification Systems," in *Interspeech*, 2019, pp. 2903–2907.

[13] V. Mingote, A. Miguel, A. Ortega, and E. Lleida, "Log-Likelihood-Ratio Cost Function as Objective Loss for Speaker Verification Systems," in *Interspeech*, 2021, pp. 2361–2365.

[14] E. Shriberg, S. Kajarekar, and N. Scheffer, "Does session variability compensation in speaker recognition model intrinsic variation under mismatched conditions?" in *Interspeech*, 2009.

[15] S. J. Wenndt and R. L. Mitchell, "Machine recognition vs human recognition of voices," in *ICASSP*, 2012, pp. 4245–4248.

[16] D. Van Lancker and J. Kreiman, "Voice discrimination and recognition are separate abilities," *Neuropsychologia*, vol. 25, no. 5, pp. 829–834, 1987.

[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," *J. Acoust. Soc. Am.*, vol. 144, no. 1, pp. 375–386, 2018.

[18] S. J. Park, A. Afshan, J. Kreiman, G. Yeung, and A. Alwan, "Target and non-target speaker discrimination by humans and machines," in *ICASSP*, May 2019, pp. 6326–6330.

[19] A. Afshan, J. Kreiman, and A. Alwan, "Speaker discrimination in humans and machines: Effects of speaking style variability," in *Interspeech*, 2020.

[20] N. Lavan, L. F. K. Burston, and L. Garrido, "How many voices did you hear? Natural variability disrupts identity perception from unfamiliar voices," *Br. J. Psychol.*, vol. 110, no. 3, pp. 576–593, 2019.

[21] J. Johnson, C. McGettigan, and N. Lavan, "Comparing unfamiliar voice and face identity perception using identity sorting tasks," *Q J Exp Psychol*, vol. 73, no. 10, pp. 1537–1545, 2020.

[22] A. Afshan, J. Kreiman, and A. Alwan, "Speaker discrimination performance for "easy" versus "hard" voices in style-matched and -mismatched speech," *J. Acoust. Soc. Am.*, vol. 151, no. 2, pp. 1393–1403, 2022.

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

[24] A. Afshan, J. Guo, S. J. Park, V. Ravi, A. McCree, and A. Alwan, "Variable frame rate-based data augmentation to handle speaking-style variability for automatic speaker verification," in *Interspeech*, 2020, pp. 4318–4322.

[25] A. Afshan and A. Alwan, "Attention-based conditioning methods using variable frame rate for style-robust speaker verification," *Interspeech*, 2022.

[26] A. Afshan, "Speaking style variability in speaker discrimination by humans and machines," Ph.D. Dissertation, University of California, Los Angeles, CA, 2022.

[27] Q. Zhu and A. Alwan, "On the use of variable frame rate analysis in speech recognition," in *ICASSP*, 2000, pp. 1783–1786.

[28] V. Ravi, S. J. Park, A. Afshan, and A. Alwan, "Voice Quality and Between-Frame Entropy for Sleepiness Estimation." in *Interspeech*, 2019, pp. 2408–2412.

[29] D. A. Van Leeuwen and N. Brummer, "An introduction to application-independent evaluation of speaker recognition systems," in *Speaker Classification I: Fundamentals, Features, and Methods*, 2007, pp. 330–353.

[30] A. Paszke *et al.*, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," *arXiv:1912.01703*, 2019.

[31] D. Povey and others, "The Kaldi speech recognition toolkit," IEEE SPS, Tech. Rep., 2011.

[32] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv:1412.6980 [cs]*, 2017.

[33] P. Keating, J. Kreiman, and A. Alwan, "A new speech database for within- and between-speaker variability," in *ICPhS XIX*, vol. 126, 2019, pp. 736–739.

[34] J. Kreiman, S. J. Park, P. A. Keating, and A. Alwan, "The relationship between acoustic and perceived intraspeaker variability in voice quality," in *Interspeech*, 2015, pp. 2357–2360.

[35] P. Keating, J. Kreiman, A. Alwan, A. Chong, and Y. Lee, "UCLA speaker variability database," 2021. [Online]. Available: https://catalog.ldc.upenn.edu/LDC2021S09

[36] ——, "UCLA speaker variability database," 2021. [Online]. Available: http://www.seas.ucla.edu/spapl/shareware.html#Data

[37] T. Hasan, R. Saeidi, J. H. Hansen, and D. A. Van Leeuwen, "Duration mismatch compensation for i-vector based speaker recognition systems," in *ICASSP*, 2013, pp. 7663–7667.

[38] M. McLaren, L. Ferrer, D. Castan, and A. Lawson, "The Speakers in the Wild (SITW) speaker recognition database." in *Interspeech*, 2016, pp. 818–822.

[39] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep Speaker Recognition," *Interspeech 2018*, pp. 1086–1090, 2018.

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

[41] 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." in *Interspeech*, 2016, pp. 2219–2222.

[42] V. Ravi, R. Fan, A. Afshan, H. Lu, and A. Alwan, "Exploring the Use of an Unsupervised Autoregressive Model as a Shared Encoder for Text-Dependent Speaker Verification," in *Interspeech*, 2020.

[43] T. Zhou, Y. Zhao, and J. Wu, "ResNeXt and Res2Net Structures for Speaker Verification," in *SLT*, 2021, pp. 301–307.

[44] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Interspeech*, 2020, pp. 3830–3834.