Towards understanding speaker perception and its applications to automatic speaker recognition: effects of speaking style variability

Amber Afshan

Dept. of Electrical and Computer Engineering, University of California, Los Angeles, CA
amberafshan@g.ucla.edu

Abstract

A speaker's voice constantly varies in everyday situations such as talking to a friend, reading aloud, talking to pets, or narrating a sad incident. These speaking style changes affect the human ability to distinguish speakers based on their voice. We investigate the effects of moderate speaking style variations on human speaker discrimination performance. We also compare human performance between "same speaker" and "different speaker" trials and find that the former heavily relies on within-speaker variability and is affected by speaking style variations. Our study shows that the human and machine approaches to speaker discrimination could be different. We model the relationship between human speaker perception and speaker acoustic variability. We also aim to use this model to improve the automatic speaker verification (ASV) systems. This work systematically analyzes the effects of speaking style variability on ASV systems. We propose an entropy-based variable frame rate (VFR) technique to address style variability by performing data augmentation using style normalized speaker representations.

Index Terms: speaker recognition, speaker perception, speaking style, data augmentation, variable frame rate

1. Motivation

Speaking style varies constantly based on the context, interlocutor, emotional, psychological, and physical state [1]. Hence, a speaker's voice displays intra-speaker variability idiosyncratic to the speaker; some speakers' voices vary more than others. Voice also varies among speakers resulting in inter-speaker variability. However, little is known about perception of intra- and inter-speaker variation. Moreover, automatic speaker recognition (ASpR) approaches assume that intra-speaker variation is much smaller than inter-speaker variation resulting in degradation in performance with variations in vocal effort, speaking style, emotional state, and physical status [2, 3, 4, 5].

In a comparison of read versus pet-directed speech (characterized by exaggerated prosody), humans consistently outperform machines in both style-matched and -mismatched conditions [5] although performance degraded due to style variations in both humans and machines. They also showed that humans use perceptual strategies different than that of ASV systems in those conditions. Moreover, forensic studies have made attempts to integrate human and machine responses [6] and found that forensic experts were able to resolve pairs falsely classified by ASV systems. Hence, modeling human perceptual approaches could assist in improving machine performances.

ASV performance suffers from a mismatch between train and test conditions [7]. External factors (environmental noise, recording conditions, channel types) have been studied extensively in ASV literature [8, 9]. Factors that are directly related to the speaker also affect ASV performance [2, 3]. Speaking style variability is one such factor. However, only a few studies have focused on speaking style variability. Style mismatch effects were studied in [4, 5]. Some others address style variabilities: a joint factor analysis framework [2, 3] and curriculum-learning-based transfer learning [10]. However, these solutions require training data with all the styles occurring in testing and thus prior knowledge of the speaking style of the test utterances.

We aim to understand the relationship between speaker perception and intra- and inter-speaker variations. In other words, in daily life variability, how likely are confusions among speakers, i.e., how often does a speaker sound less like him- or herself and more like someone else? And how do listeners deal with such situations? Moreover, can we develop better algorithms using our knowledge of speaker perception?

One of the primary databases used in this work is the UCLA Speaker Variability Database [11, 12, 13]. It captures commonly occurring variations in speech from 101 female and 101 male speakers. In this work, the speech tasks from the database include reading sentences (scripted speech); narrating a recent neutral, happy, or annoying conversation (unscripted affective speech); making a telephone call to a familiar person (unscripted casual conversational speech); and talking aloud to pets in a video (pet-directed speech). This database was recorded in a sound-attenuated booth at a sampling rate of 22 kHz.

2. Perception of speaking style variability

We aim to understand speaker discrimination abilities and investigate the effects of speaking style variations (read versus conversational speech). We examined the task of unfamiliar speaker discrimination from text-independent and short (∼3s), utterances [14]. Thirty normal-hearing listeners (24 natives and 6 non-natives) were asked to decide if two samples are from the same speaker or not. The stimuli were derived from 40 female speakers from the UCLA database. They were self-reported native English speakers, verified *post hoc* by two linguists.

Results in Table 1 show a decrease in speaker discrimination performance with style: starting from the style-matched condition of read speech–read speech followed by conversation–conversation and the worst in the style-mismatched trials of read speech–conversation. Listeners were more confident when performing the "same speaker" task vs the "different speaker" task in the style-matched trials of read speech. However, the listener confidence in the "same speaker"

Table 1: *Speaker discrimination performance in terms of equal error rates (EER, %) for native and non-native listeners.*

Listeners	read-read	conv.-conv.	read-conv.
Native	6.96	15.12	20.68
Non Native	12.39	23.22	31.46

trials decreased for style-matched conversation trials and was the worst for the style-mismatched ones. "Different speaker" trials did not follow this pattern. Thus, the effects of moderate speaking style variability are higher on the performance of the "same speaker" task [14]. In general, native English listeners were more confident and more reliable than non-natives.

2.1. Relationship between speaker perception and speaker acoustic variability

Lee et al. [15, 16] showed that within-speaker acoustic variability shares a similar structure with group acoustics. However, the majority of the within-speaker variability is idiosyncratic. Our current work [17] models the relationship between speaker perception and speaker acoustic variability. We found that humans rely on the distances between speakers in the shared structure when "telling speakers apart." Whereas, they relied on speaker-specific idiosyncrasies when "telling speakers together."

2.2. Relationship between human and machine speaker discrimination performance

We compare human and machine performance in speaker discrimination tasks for read versus conversational speech. Humans performed better than machines in style-matched conditions. Unlike humans, machines showed a similar pattern in reliability irrespective of the speaking style in the "same speaker" and "different speaker" tasks [14], suggesting that there may be no specific difference between the two decisions in machines. Speakers who populate the subsets that humans and machines found easy or difficult to distinguish were not similar. Taken together these results, suggest that ASV systems use a different strategy than humans for speaker discrimination. Moreover, humans use different approaches based on the task ("same speaker" or "different speaker"), whereas machines rely on the same strategies irrespective of the tasks.

3. Speaking style variability in automatic speaker verification

We analyze the performance of x-vector [8]/PLDA (probabilistic linear discriminant analysis; [18]) system during style-matched and -mismatched conditions. The latter resulted in performance degradation as shown in Table 2. Data augmentation is generally performed to address data mismatch using either a larger database with all the different conditions per speaker or artificially generates data for augmentation. However, in this case these approaches are not feasible. First, there is no large publicly available data set with multiple styles per speaker. Second, artificially generating style variations is an active area of research [19] and hence might not deliver desired results. Thus, we propose using entropy-based VFR [20] technique for data augmentation by generating style-normalized variants [21].

Speaking style variability could result in differences in many aspects of the speech signal. For example, change from read to conversational speech include variation in speaking rate and inconsistent pauses between words. Moreover, conversations often result in vowels being modified or reduced in duration and inconsistencies in the release of word-final plosive bursts [22]. Similar differences occur among other speaking styles [23]. Inter-frame entropy is high with a rapid change of spectral characteristics from a high speech rate and/or short pause. However, inter-frame entropy becomes lower with a slower speech rate. These changes partially reflect the variations caused by speaking style. Thus, the proposed VFR technique uses inter-frame entropy to dynamically change frame rate and obtain a uniform entropy across frames, resulting in partially style-normalized speaker representations. Given the small size of the UCLA database, we restricted the augmentation experiments to PLDA adaptation configurations. Table 2 provides x-vector performance degraded drastically in style-mismatched conditions when compared to style-matched ones. The proposed approach significantly improved ASV performance in such conditions. The proposed approach was comparable to the best-case scenario of multi-style training. Thus, the VFR technique could be effective for applications with a lack of multi-style training data and also when there is no prior knowledge about testing conditions.

Table 2: *ASV results on baseline, VFR normalized augmentation (proposed) and multi-style augmentation (best-case) are presented 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
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

4. Future directions

4.1. Applications of perception model to automatic speaker verification

This work raises the question of whether we could implement human perceptual strategies to improve ASV performance. Analysis of acoustic variability suggested that the listeners' find it easier to "tell speakers together" when they rely on speaker-specific idiosyncrasies while they "tell speakers apart" based on their distances within a shared acoustic space. In the next step, we aim to extend this knowledge and use particular features that humans relied on for the "same speaker" and "different speaker" tasks to improve recognition in the face of speaking style variability, a task that listeners routinely perform with ease.

4.2. Style-robust automatic speaker verification systems

From our previous work on the entropy-based VFR technique for data augmentation, we know that VFR can generate partially style-normalized speaker representations. As an extension, we aim to develop speaking style-robust models that use VFR in their model structure, resulting in style-normalized speaker representations. For this, we plan to utilize the VFR technique in conjunction with self-attention [24] for building ASV systems.

5. Acknowledgments

The author would like to thank their advisor, Prof. Abeer Alwan, for her invaluable guidance and support in this work. They would also to thank Prof. Jody Kreiman and Dr. Alan McCree for their guidance. This work was supported in part by the NSF.

6. References

[1] J. Kreiman and D. Sidtis, *Foundations of voice studies: An interdisciplinary approach to voice production and perception*. WileyBlackwell, Walden, MA: John Wiley & Sons, 2011, pp. 245-246.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] V. Hughes, P. Harrison, P. Foulkes, P. French, C. Kavanagh, and E. S. Segundo, "Mapping Across Feature Spaces in Forensic Voice Comparison: The Contribution of Auditory-Based Voice Quality to (Semi-)Automatic System Testing," in *Interspeech 2017*. ISCA: ISCA, Aug. 2017, pp. 3892–3896.

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

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

[9] Y. Zhu, T. Ko, D. Snyder, B. Mak, and D. Povey, "Self-Attentive Speaker Embeddings for Text-Independent Speaker Verification," in *Interspeech 2018*. ISCA, Sep. 2018, pp. 3573–3577.

[10] 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.

[11] 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.

[12] P. Keating, J. Kreiman, A. Alwan, A. Chong, and Y. Lee, "UCLA speaker variability database," 2021, (Last viewed July 28, 2021). [Online]. Available: http://www.seas.ucla.edu/spapl/shareware.html#Data

[13] 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.

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

[15] Y. Lee, P. Keating, and J. Kreiman, "Acoustic voice variation within and between speakers," *The Journal of the Acoustical Society of America*, vol. 146, no. 3, pp. 1568–1579, Sep. 2019.

[16] Y. Lee and J. Kreiman, "Variation in voice quality within speakers," *The Journal of the Acoustical Society of America*, vol. 145, pp. 1930–1930, May 2019.

[17] A. Afshan, J. Kreiman, and A. Alwan, "Speaker discrimination for "easy" versus "hard" voices in style-matched and -mismatched speech," manuscript submitted for publication.

[18] P. Kenny, T. Stafylakis, P. Ouellet, M. J. Alam, and P. Dumouchel, "PLDA for Speaker Verification with Utterances of Arbitrary Duration," in *Proc. ICASSP*, 2013, pp. 7649–7653, iSSN: 15206149.

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

[20] 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.

[21] 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.

[22] 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.

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

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is All you Need," *Advances in neural information processing systems*, pp. 5998–6008, 2017.