DOMAIN ADAPTATION FOR SPEAKER RECOGNITION IN SINGING AND SPOKEN VOICE

Anurag Chowdhury[*], *Austin Cozzo*[*†], *Arun Ross*[*]

[*]Michigan State University, [†]Rank One Computing Corporation

chowdh51@msu.edu, austin.cozzo@rankone.io, rossarun@msu.edu

ABSTRACT

In this work, we study the effect of speaking style and audio condition variability between the spoken and singing voice on speaker recognition performance. Furthermore, we also explore the utility of domain adaptation for bridging the gap between multiple speaking styles (singing versus spoken) and improving overall speaker recognition performance. In that regard, we first extend a publicly available singing voice dataset, JukeBox, with corresponding spoken voice data and refer to it as JukeBox-V2. Next, we use domain adaptation for developing a speaker recognition method robust to varying speaking styles and audio conditions. Finally, we analyze the speech embeddings of domain-adapted models to explain their generalizability across varying speaking styles and audio conditions.

Index Terms— Speaker Recognition, Singing Voice, Speaking style, Domain Adaptation, Deep Learning

1. INTRODUCTION

Speaker recognition, or voice biometrics, entails comparing two speech samples to determine if the same individual produced them. Most speaker recognition systems assume "ideal audio conditions," such as minimal background noise, neutral speaking style, and normal vocal effort for optimal performance [1]. However, such an assumption is an oversimplification of practical voice biometrics scenarios. While several recently developed methods have focused on performing speaker recognition in the presence of background noise and degradations, the majority of them only consider spoken voice (i.e., speech uttered in a neutral speaking style) for training and evaluating their approaches [1, 2]. Spoken voice, however, only represents a limited range of possible vocal dynamics for a speaker [3]. Therefore, methods based on neutral spoken voice suffer performance degradation with varying speaker style and effort [4, 5].

Among possible speaking styles, singing voice presents a particularly less explored mode of speaker recognition [6]. The challenges of singer recognition – speaker recognition where the speaking style is singing – differs from traditional speaker recognition due to the much broader range of perceptual qualities and underlying physiological dynamics apparent in the singing voice [7, 8, 9]. The singing voice's features are further diversified by the singing style, which is influenced by the genre and accompanying music [10]. Singing voice, thus, serves as an example for a wide variety of speaking styles and audio conditions that present a challenge to traditional speaker recognition systems [11, 12].

A previous work assembled a singing voice dataset, JukeBox [5], and demonstrated the challenges of performing singer recognition using models pre-trained on spoken voice. Furthermore, the pre-trained models were fine-tuned using singing voice to improve singer verification performance. However, the fine-tuned models were not evaluated on spoken voice to determine their generalizability across different speaking styles. In addition, the original JukeBox dataset does not contain any spoken voice samples corresponding to a person's singing voice, limiting its utility for cross-domain speaker verification, i.e., matching a person's singing voice to their spoken voice or vice versa.

Following these observations, the contributions of this work are as follows. We first extend the original JukeBox dataset to include spoken voice samples for the subjects in the test set of the JukeBox dataset, now referred to as JukeBox-V1. This extended dataset, referred to as JukeBox-V2, enables the evaluation of speaker recognition methods across varying speaking styles and which we make publically available.[1] We next apply domain adaptation (DA) to equitably learn discriminative features for both the singing and spoken voice, reaffirming the utility of DA when presented with the novel intrinsic variability in the speaking style demonstrated by the singing voice. Finally, we also analyze the impact of speaking style variability between the singing and spoken voices on the learned feature space.

2. MOTIVATION

The original study of the JukeBox-V1 dataset *fine-tuned* the pre-trained speaker recognition models using singing voice for improving singer recognition performance [5]. However, as we will demonstrate in this paper, these fine-tuned models result in performance degradation on spoken voice (see Section 5 and Fig. 2). Therefore, the performance degradation on *spoken* voice accompanied by the modest increase in performance on *singing* voice indicates fine-tuning as a sub-optimal solution and the problem needs to be carefully revisited.

This performance disparity suggests the presence of a **domain gap** between the speaking styles, as also alluded to by the noticeable perceptual differences between the spoken and

[1] http://iprobe.cse.msu.edu/datasets/jukebox_v2.html

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

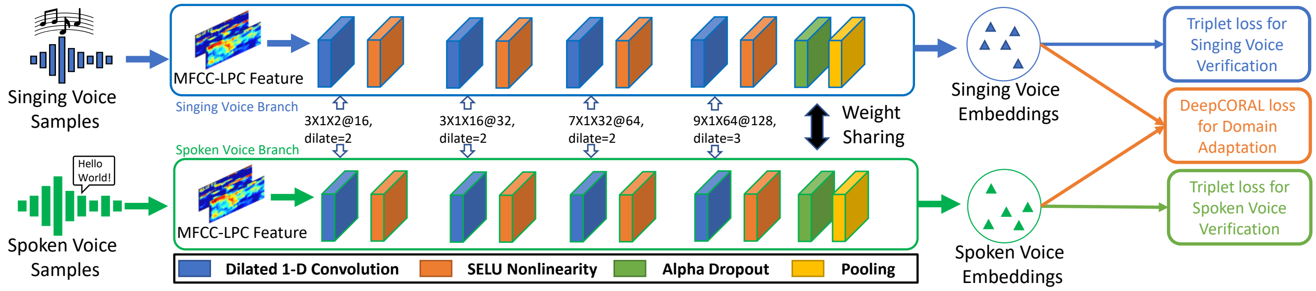

Fig. 1: A visual representation of the domain-adaptation-based 1D- CNN framework proposed in Section 3.

the singing voice. Similar to the singing voice, the whispered voice is a relatively well-studied speaking style [13, 14] that contains different acoustic characteristics to the spoken voice. For example, the F1 and F2 formants of the whispering [13] and singing [3] voices deviate from the spoken voice. Due to such intrinsic variations in the acoustic characteristics, the whispering voice is often treated as a speaking style variation characterized by a low vocal effort and unvoiced speech [13, 14]. Similarly, as done in [5, 11] and in this work, the singing voice is assumed to be a speaking style variation characterized by an increased vocal range.

Speaker recognition systems are often adversely affected by a wide variety of perturbations, both extrinsic and intrinsic. While extrinsic perturbations such as channel variability are often addressed by techniques such as dataset variability compensation [15], intrinsic perturbations such as language and vocal effort variability are often resolved using DA [16, 17]. Therefore, in this work we use unsupervised DA to develop speaker recognition methods robust to speaking style variability. Specifically, we use the CORAL [18], CORAL+ [16] and DeepCORAL [19] techniques due to their simplicity and demonstrated effectiveness for bridging the domain gap created by intrinsic variabilities in the human voice [16]. To the best of our knowledge, **this is the first work to explore DA for developing speaker recognition models robust to speaking style variabilities**.

3. DOMAIN ADAPTATION-BASED SPEAKER RECOGNITION FRAMEWORK

Speaker-dependent speech features such as phoneme duration, mean fundamental frequency (F0), and formant center frequencies that are crucial for speaker modeling differ vastly between the speaking and the singing voice, thus creating a domain gap between the two speaking styles [3, 13]. In this work, therefore, we extend the fine-tuned models used in [5] to DA-based approaches. The 1D-Triplet-CNN [2], the best performing model in [5], is adapted to include DeepCORAL loss [18, 19] as shown in Fig. 1. Additionally, in order to evaluate the effectiveness of DA on speaking style, we also evaluate the other models used in [5]. The probabilistic linear discriminant analysis (PLDA) classifiers iVector-PLDA [20] and xVector-PLDA [21] are combined with the CORAL+ algorithm [16] for an analogous addition of DA. The use of both PLDA- and deep learning-based approaches allows this work to more generally demonstrate the efficacy of DA in both classical and state-of-the-art speaker recognition models.

The proposed 1D-CNN framework (Fig. 1) consists of two identical 1D-CNN [2] branches with shared weights. Each branch extracts an MFCC-LPC feature path [2] which is then used in the adaptive triplet mining approach outlined in [22]. To retain comparability with [5], the same triplet construction method is used for both spoken and singing voice samples. The two set of triplets from the singing and spoken voice data are then used to minimize the corresponding cosine triplet embedding losses [2], L_{si} and L_{sp} respectively, for training the 1D-CNN branches. The functional form of both the losses is given by:

$$L(S_a, S_p, S_n) = \sum_{a,p,n}^{N} \cos(g(S_a), g(S_n)) \qquad (1)$$
$$- \cos(g(S_a), g(S_p)) + \alpha_{margin}$$

Here, N is the total number of triplets drawn by the adaptive triplet mining method [22]. α_{margin} is the margin of the minimum distance between positive and negative samples and is a user-tunable hyper-parameter.

For performing DA between the singing and spoken voice samples, we minimize the distance between the covariances C_{si} and C_{sp}, constituting the DeepCORAL loss [18, 19], of the singing and spoken voice embeddings $g(x_{si})$ and $g(x_{sp})$. The DA loss (L_{DA}) is given by:

$$L_{DA}(g(x_{si}), g(x_{sp})) = \frac{1}{4d^2} \left\|(C_{si} - C_{sp})\right\|_F^2 \qquad (2)$$

Here, $\|\cdot\|_F^2$ is the squared matrix Frobenius norm and d is the dimensionality of the voice embeddings. It is worth noting that DeepCORAL loss has no fixed-direction from one domain to another and would be symmetrical if C_{si} and C_{sp} were swapped. The combined loss L for the entire framework is given as follows:

$$L = \alpha_1 L_{si} + \alpha_2 L_{sp} + \beta L_{DA} \qquad (3)$$

Here, α_1, α_2, and β are user-tunable hyper-parameters (in our experiments, 1, 10, and 10, respectively) that control the effect of individual losses on the combined loss. For evaluation, speech embeddings extracted using the trained model are matched using the cosine similarity metric.

For applying CORAL to the PLDA based approaches such as iVectorPLDA and xVectorPLDA, we use the CORAL+ algorithm for performing unsupervised domain adaptation of the trained PLDA classifiers [16].

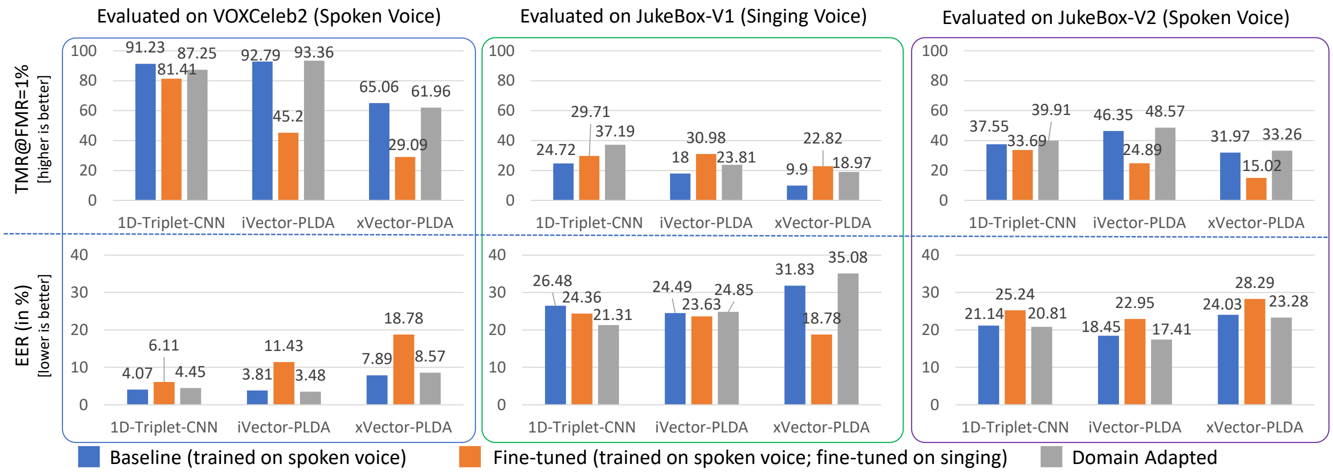

Fig. 2: Summary of verification performance (Top: TMR@FMR=1%, Bottom: EER (in%)) across different evaluation conditions. Note the increase in *singer* recognition performance in both fine-tuned (orange bars) and domain adapted (grey bars) models and increase in speaker recognition performance in domain adaptation over fine-tuning.

4. DATASETS AND EXPERIMENTAL PROTOCOLS

4.1. Datasets

In this work, we use the VoxCeleb2 [23] dataset to train and evaluate the speaker recognition models on spoken voice data (i.e., spoken-to-spoken scenarios). Continuing the evaluation scenario in [5], we use the same subset of $5,994$ video samples corresponding to the $5,994$ celebrities in the VoxCeleb2 dataset to form the training set. While more speaking voice data would improve performance on spoken voice evaluations (particularly in the case of xVector-PLDA), similarly sized datasets for both the spoken and singing voice domains are important for performing effective fine-tuning/DA and maintaining a fair comparison to [5]. A random subset of 118 video samples corresponding to 118 celebrities forms the test set of VoxCeleb2 dataset. Speech from each video sample is split into multiple non-overlapping 5-second long audios.

We use the JukeBox-V1 dataset [5] to fine-tune, train for DA, and evaluate the speaker recognition models on singing voice. Training data is augmented by splitting each song into multiple non-overlapping 30-second long segments, as in [5]. We use the test set of Jukebox-V1 for singing-to-singing evaluation. Furthermore, we collected four 5-second long spoken voice samples corresponding to 92 out of 98 subjects in JukeBox-V1's test set for a total of 368 samples. We could not locate any spoken voice data for the remaining six subjects. We collected the data by identifying interviews of each singer on YouTube and manually isolating the target's speech audio. This extension of the test set of the JukeBox-V1 dataset, referred to as JukeBox-V2, enables future cross-domain evaluation of speaker recognition algorithms.

4.2. Experiments Performed

In order to demonstrate the effectiveness of DA on speaking style variability, we create three categories of models: 1) Baseline models, trained only on spoken voice data (VoxCeleb2 subset); 2) Fine-tuned models, trained on spoken voice data (VoxCeleb2 subset) and fine-tuned on singing voice data (JukeBox-V1 training set); and 3) DA-based models, trained on both spoken and singing voice data (VoxCeleb2 subset & JukeBox-V1 training set). These models are grouped in Fig. 2 by color, with blue, orange, and grey representing Baseline, Fine-tuned and DA-based models, respectively. Each category contains the same 3 models as [5], iVector-PLDA [20], xVector-PLDA [21], and 1D-Triplet-CNN [2], to evaluate the imapct of DA on classical and state-of-the-art speaker recognition models.

The three categories are evaluated on three different test sets: 1) VoxCeleb2's test set, 2) JukeBox-V1's test set, and 3) JukeBox-V2. The VoxCeleb2 test allows for large-scale spoken-to-spoken evaluation, the JukeBox test set allows for singing-to-singing evaluation, and the JukeBox-V2 set allows for spoken-to-spoken evaluation on the same identities as the singing-to-singing of the JukeBox-V1 test set.

For performance metrics, we use True Match Rate at a False Match Rate (TMR@FMR) of 1% and Equal Error Rate (EER in $\%$). We choose EER as a standard metric for speech processing and TMR@FMR=1% in accordance with NIST's criteria for biometric performance comparison [24].

5. RESULTS

All the *baseline* models attain significantly lower performance on singing voice than their spoken voice counterpart in the JukeBox-V2 dataset. This reinforces the challenges faced by models trained on spoken voice alone when evaluated on the singing voice [5]. As also noted in [5], the xVector-PLDA model's relatively lower performance is likely attributable to the training data being insufficient for training xVector's considerably larger parameter space (4.2M) compared to the 89K parameters in the 1D-Triplet-CNN and an even smaller parameter space in the iVector-PLDA model. On average, the *fine-tuned* models, compared to the baseline models, demonstrate an increase in performance (TMR@FMR=1%) on singing voice by $\sim 10\%$, but they also demonstrate an average

Table 1: Poor speaker verification results on cross-modal voice data from the JukeBox-V2 dataset justifying the application of DA (see Fig. 2)

Models	Category	TMR@FMR=1%	EER (in %)
Baseline	1D-Triplet-CNN	1.39	48.17
	iVector-PLDA	2.37	49.11
	xVector-PLDA	0	48.14
Fine-tuned	1D-Triplet-CNN	0.6	43.02
	iVector-PLDA	1.36	43.54
	xVector-PLDA	1.2	44.82
DA-Based	1D-Triplet-CNN	1.67	42.11
	iVector-PLDA	1.73	44.64
	xVector-PLDA	1.58	42.78

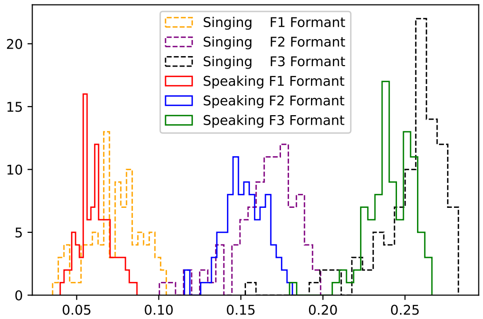

Fig. 3: Histogram plots of the first three formants (F1-F3) of spoken and singing speech from the JukeBox-V2 dataset

performance *loss* on spoken voice alone in the JukeBox-V2 and the VoxCeleb2 datasets by ∼14% and ∼31%, respectively. This demonstrates the fine-tuned model's lack of generalizability across speaking styles. **The *domain-adapted models* outperform** (TMR@FMR=1%) their corresponding baseline models on singing and spoken voices from the JukeBox-V2 dataset by an average of ∼9% and ∼2%, respectively. DA also reduces the average performance loss on spoken voice in the VoxCeleb2 dataset from ∼31% to ∼2%. Furthermore, in the spoken voice JukeBox-V2 dataset, the average performance loss of ∼14% is converted to a performance *gain* of ∼2%. **This demonstrates generalizability of the domain-adapted models across speaking styles.** However, in the cross-domain speaker recognition experiments in Table 1, the DA models do not offer any significant performance improvement over the baseline methods, as they are unable to map a person's spoken voice to their singing voice. Therefore, we believe more cross-domain training data (currently unavailable) is essential to learn the mapping between an individual's singing and spoken voice.

6. ANALYSIS

In Section 5, we experimentally demonstrated the domain gap between the singing and spoken voice. A histogram of the first three formants of the singing and spoken voice in the JukeBox-V2 dataset qualitatively verify the presence of the domain-gap in Figure 3. In this section, we inspect the effect of this domain gap in the feature space learned by the 1D-Triplet-CNN-, iVector-PLDA-, and xVector-PLDA-based baseline speaker recognition models, trained on spoken voice alone. In comparison, we analyze the feature space learned by the different approaches when combined with DA to understand the effect of DA across the two speaking styles. To this end, we compare the t-SNE [25] plots of speech embeddings of spoken and singing voice data from the JukeBox-V2's test set, extracted by the 1D-Triplet-CNN, iVector-PLDA, and xVector-PLDA-based models, both without and with DA (Fig. 4). In Fig. 4, the singing and spoken voice embeddings extracted by the models without DA form separate perceived clusters demonstrating the presence of the domain gap. **However, in the models with DA, the overlap in perceived clusters increases notably and indicates that the domain gap is reduced.** This difference in the speech embedding clusters between the DA and non-DA models we hypothesize is a direct effect of the CORAL+ and DeepCORAL loss used for performing DA. The DA minimizes the covariance between the speech embeddings of the two speaking styles, thereby merging their clusters and partially bridging the domain gap.

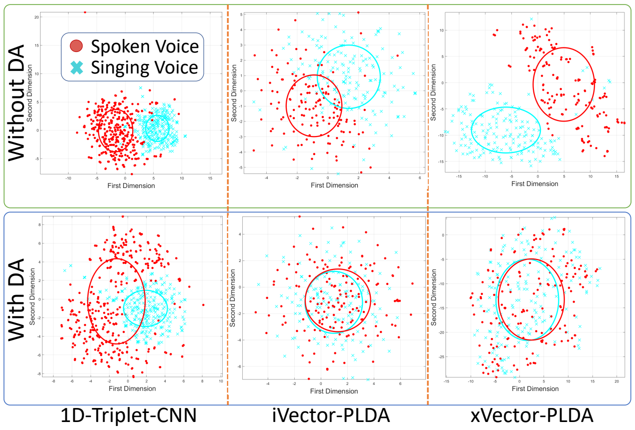

Fig. 4: t-SNE plots of the speech embeddings (with and without DA) of singing and spoken voice from the JukeBox-V2 dataset. The circles were added to indicate the apparent cluster boundaries. After DA, the apparent domain gap is reduced leading to overlap of the circle as shown in the lower row. The without DA methods refer to the baseline models.

7. SUMMARY AND FUTURE WORK

Singing voice data introduces the challenges of varying speaking style and background noise to speaker recognition. Therefore, training speaker recognition models using domain adaptation, as evidenced by the experiments, has the potential to improve their generalizability between spoken and singing voices. We also assembled the JukeBox-V2 dataset to demonstrate the challenges of cross-domain speaker recognition between spoken and singing voices. Toward that end, it may be valuable to explore the benefits of combining speaking style-specific speech filter banks [26] with DA to improve cross-domain speaker recognition performance. Additionally, the availability of cross-domain training data across varied speaking styles is important to develop cross-domain speaker recognition systems.

8. REFERENCES

[1] Tomi Kinnunen and Haizhou Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, 2010.

[2] Anurag Chowdhury and Arun Ross, "Fusing MFCC and LPC features using 1D Triplet CNN for speaker recognition in severely degraded audio signals," *IEEE Transactions on Information Forensics and Security*, 2020.

[3] John HL Hansen, Marigona Bokshi, and Soheil Khorram, "Speech variability: A cross-language study on acoustic variations of speaking versus untrained singing," *The Journal of the Acoustical Society of America*, 2020.

[4] Elizabeth Shriberg et al., "Effects of vocal effort and speaking style on text-independent speaker verification," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[5] Anurag Chowdhury, Austin Cozzo, and Arun Ross, "Jukebox: A multilingual singer recognition dataset," *INTERSPEECH*, 2020.

[6] Mahnoosh Mehrabani and John Hansen, "Speaker clustering for a mixture of singing and reading," *INTERSPEECH*, 2012.

[7] William S Brown Jr, Howard B Rothman, and Christine M Sapienza, "Perceptual and acoustic study of professionally trained versus untrained voices," *Journal of Voice*, 2000.

[8] Ray Daniloff et al., "Allophonic variation in spoken and sung speech," *The Journal of the Acoustical Society of America*, 1994.

[9] William S Brown Jr, Elizabeth Hunt, and William N Williams, "Physiological differences between the trained and untrained speaking and singing voice," *Journal of Voice*, 1988.

[10] Johan Sundberg, "The acoustics of the singing voice," *Scientific American*, 1977.

[11] Mahnoosh Mehrabani and John HL Hansen, "Singing speaker clustering based on subspace learning in the GMM mean supervector space," *Speech Communication*, 2013.

[12] Shichao Hu et al., "Large-scale singer recognition using deep metric learning: an experimental study," in *International Joint Conference on Neural Networks*. IEEE, 2021.

[13] Taisuke Ito, Kazuya Takeda, and Fumitada Itakura, "Analysis and recognition of whispered speech," *Speech Communication*, 2004.

[14] Xing Fan and John Hansen, "Speaker identification within whispered speech audio streams," *IEEE Transactions on Audio, Speech, and Language Processing*, 2011.

[15] Hagai Aronowitz, "Inter dataset variability compensation for speaker recognition," in *IEEE ICASSP*, 2014.

[16] Kong Aik Lee, Qiongqiong Wang, and Takafumi Koshinaka, "The CORAL+ algorithm for unsupervised domain adaptation of PLDA," in *IEEE ICASSP*, 2019.

[17] Fahimeh Bahmaninezhad, Chunlei Zhang, and John HL Hansen, "An investigation of domain adaptation in speaker embedding space for speaker recognition," *Speech Communication*, 2021.

[18] Baochen Sun, Jiashi Feng, and Kate Saenko, "Return of frustratingly easy domain adaptation," in *AAAI Conference on Artificial Intelligence*, 2016.

[19] Baochen Sun and Kate Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *European Conference on Computer Vision*. Springer, 2016.

[20] Najim Dehak et al., "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, 2011.

[21] David Snyder et al., "X-Vectors: robust DNN embeddings for speaker recognition," in *International Conference on Acoustics, Speech, and Signal Processing*, 2018.

[22] Anurag Chowdhury and Arun Ross, "DeepVOX: Discovering features from raw audio for speaker recognition in degraded audio signals," *arXiv preprint arXiv:2008.11668*, 2020.

[23] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "Voxceleb2: Deep speaker recognition," in *INTERSPEECH*, 2018.

[24] Craig S Greenberg et al., "Two decades of speaker recognition evaluation at the national institute of standards and technology," *Computer Speech & Language*, 2020.

[25] Laurens van der Maaten and Geoffrey Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, 2008.

[26] Chi Zhang and John HL Hansen, "Whisper-island detection based on unsupervised segmentation with entropy-based speech feature processing," *IEEE Transactions on Audio, Speech, and Language Processing*, 2010.