



A Siamese Network-Based Framework for Voice Mimicry Proficiency Assessment Using X-Vector Embeddings

Bhasi K.C.^{1,3}, Rajeev Rajan^{2,3}

¹Government Engineering College Barton Hill, India

²Government Engineering College Idukki, India

³APJ Abdul Kalam Technological University, India

bhasikc@gmail.com, rajeev@gecidukki.ac.in

Abstract

This paper evaluates the quality of mimicked speech by computing X-vector based speaker embeddings derived from spectral and prosodic features. These embeddings are processed through a long short-term memory (LSTM)-based Siamese Network. The best mimicry artist is first identified through a perception test, and we then examine whether the Siamese Network predicts the same artist. A correct prediction is counted as a hit when the model assigns the highest similarity score (rank-1) to the artist identified by the mean opinion score (MOS). Performance evaluation is conducted on the MIMICz dataset using Top-K hit rate. The results highlight the effectiveness of score-level fusion of spectral and prosodic speaker embeddings in assessing voice mimicry competency, with the fusion approach achieving a Top-1 hit rate of 72% and a Top-2 hit rate of 88%.

Index Terms: X-Vector, Siamese Network, perception test, Top-K hit rate, score-level fusion

1. Introduction

Mimicry involves the replication of distinct characteristics across various domains, including public figures, entertainers, singers, and even sounds from nature, such as bird calls and animal vocalizations. It is commonly observed in comedic performances, stage acts, and even in the animal kingdom, where it serves as a tool for communication or self-defense. Voice mimicry, or vocal impersonation, refers to the ability of an individual to imitate another person's speech patterns and vocal traits [1]. Modifying one's voice can serve multiple purposes, such as concealing the speaker's identity or convincingly mimicking someone else, often achieving both simultaneously [1]. Producing an authentic voice imitation requires precise control of articulatory mechanisms, demanding coordinated movement of the lips, tongue, and other speech structures.

The proposed system introduces an automated framework to evaluate voice mimicry by identifying the most proficient artist based on vocal resemblance. Proficiency, in this context, is defined as the degree to which an impersonator can mimic a target voice with maximum similarity. This assessment is conducted through a systematic analysis of both spectral and prosodic features, quantifying the similarity between mimicked and original speech. Prosodic attributes are essential in voice imitation, as they capture a speaker's unique vocal signature and emotional nuances [2]. Research shows that incorporating prosodic elements into filter-based recognition frameworks enhances performance by providing additional distinguishing features [3]. Effective mimicry requires precise pitch modulation, controlled speech rate variations, and strategic loudness adjustments to replicate the intricate details of a target voice [4].

Mimicked speech quality assessment has gained attention

due to its impact on speaker verification, particularly in spoofing attacks. Mel frequency cepstral coefficients (MFCCs) have been widely used for evaluating text-independent mimicked speech [5, 6], with GMM-based classifiers utilizing log-likelihood scores to determine authenticity. Prosodic features, including pitch, jitter, and shimmer, have also been explored for assessing similarity between mimicked and original speech [7], with dynamic time warping improving temporal alignment [8]. Studies on voice disguise reveal that certain speaker-specific traits, such as phoneme locus, remain challenging to replicate even for skilled impersonators [9]. The i-vector framework, integrating MFCC-based i-vectors with phonation features, has been employed to identify the most proficient imitators [10]. GAN-based models learn spectrographic representations of source and target voices to facilitate mimicry, emphasizing the role of spectral content [11]. A Sequence-to-Sequence Attentional Siamese Neural Network (Seq2Seq-ASNN) enhances speaker verification by aligning evaluation and enrollment utterances, outperforming traditional i-vector and d-vector approaches [12]. Speech morphing techniques modify spectral and prosodic characteristics to transform a source speaker's voice into that of a target speaker, achieving more natural-sounding impersonations [13].

X-vectors have been extensively applied in speaker recognition tasks [14, 15, 16], offering robust discriminative features that enhance voice mimicry evaluation. Their integration into mimicry detection systems improves performance, but recent advancements in deep learning have led to the exploration of alternative feature extraction techniques. Furthermore, X-vector-based speaker anonymization alters a speaker's voice to resemble a pseudo-speaker, with its effectiveness influenced by factors such as speaker distance, selection criteria, gender, and consistency across utterances [17].

2. Proposed System

Our system depicted in Figure 1 employs a hierarchical approach to assess the similarity between original and mimicked voices. First, spectral and prosodic features are extracted from the audio recordings of both training and testing (mimicked) samples. These extracted features are then used to generate speaker embeddings in the form of X-vectors through a time-delay neural network (TDNN). The obtained embeddings are subsequently processed by a Siamese Network, which computes a similarity score between the original and mimicked voices. By evaluating the similarity scores across all mimicry artists, we establish a ranking system, where the artist with the highest similarity score is identified as the best mimic. This framework enables a quantitative assessment of voice mimicry based on learned speaker representations.

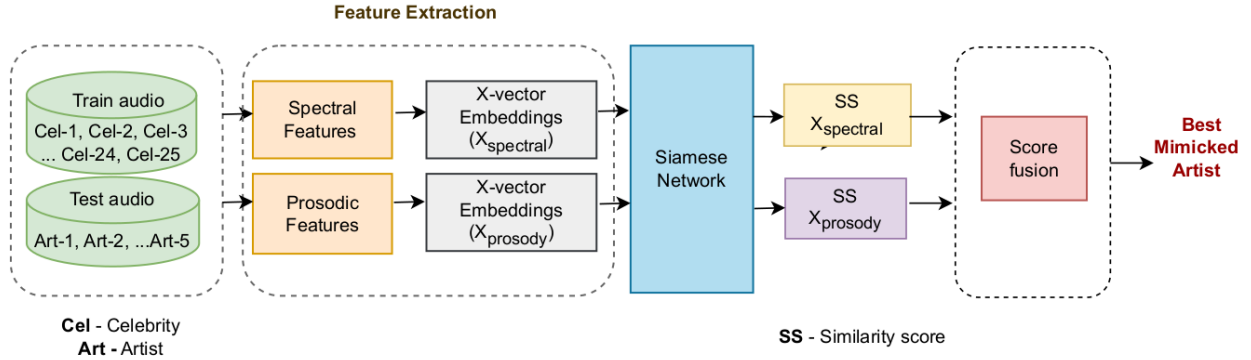


Figure 1: Block diagram of the proposed system

2.1. Feature Extraction

Spectral characteristics are derived from audio signals and encompass a range of spectral features [18] with an overall dimension of 31, including MFCCs, spectral bandwidth, spectral roll-off, spectral flux, spectral centroid and spectral contrast. Similarly, prosodic features [19], with an overall dimension of 20, include loudness, pitch contour, energy contour, duration, speaking rate, shimmer, jitter and tempogram ratio. These features are subsequently converted into X-vector embeddings of dimension 512.

The x-vector speaker embedding [20] has been successfully employed in tasks such as speaker verification [21], diarization [22], and recognition [23]. X-vectors are extracted from speech signals using a time-delay neural network (TDNN) architecture [24], which, once trained, can generate X-vectors from unseen speech samples. X-vectors are computed from both spectral and prosodic features, denoted as $X_{spectral}$ [25] and $X_{prosody}$, respectively.

Table 1: Base Network Model Summary

No.	Output Size	Description
1	(None,1, 512)	Input Layer
2	(None, 64)	LSTM Layer
3	(None, 64)	Batch Normalization
4	(None,32)	Dense Layer
5	(None,32)	Batch Normalization

Table 2: Siamese Network Model Summary

No.	Output Size	Description
1	(None,1, 512)	Input Layer A
2	(None,1, 512)	Input Layer B
3	(None,32)	Shared Embedding from Base Network A
4	(None,32)	Shared Embedding from Base Network B
5	(None,32)	Absolute Difference (Lambda)
6	(None,1)	Dense Layer (Output)

2.2. Evaluation System

The proposed system is built upon a Siamese network designed to measure similarity between voice samples. It consists of two

key components: a Base Network for feature extraction and a Siamese Network for similarity measurement.

The Base Network shown in Table 1 extracts features from voice input of shape (1,512) using an LSTM layer with 64 units and tanh activation to capture temporal dependencies. A batch normalization layer follows to stabilize training. A dense layer with 32 units and ReLU activation transforms the LSTM output into a lower-dimensional embedding, followed by another batch normalization layer for stability. The network learns a compact representation of the voice sample.

The Siamese Network shown in Table 2 is designed to compare pairs of voice samples and determine their similarity [26]. It consists of two identical instances of the Base Network, which share weights to ensure consistency in feature extraction. Given two input samples, they are processed through the Base Network to generate embeddings. The absolute difference between these embeddings is computed using a Lambda layer, and the result is passed through a Dense layer with sigmoid activation, which outputs a similarity score between 0 and 1.

The evaluation framework constructs sample pairs for training a Siamese Network [27] using a contrastive loss function. Positive pairs consist of two samples from the same class (celebrity), while negative pairs pair a target class sample with one from a different class. The model learns to minimize distances for matching pairs and maximize them for non-matching pairs. During testing, similarity scores are computed by comparing each test sample to its corresponding class sample, providing an effective metric for distinguishing authentic and mimicked voices.

3. Performance Evaluation

3.1. Dataset

The proposed system is tested on the MIMICz dataset [28], which includes celebrity audio samples from movies and entertainment programs. The imitations by five artists, labeled Artist-1 to Artist-5, are recorded in a studio at a 44.1 kHz sampling rate. The target voices consist of twenty-five renowned Malayalam¹ movie actors. This study focuses on the "text-independent" aspect to analyze vocal characteristic effects.

¹<https://en.wikipedia.org/wiki/Malayalam>

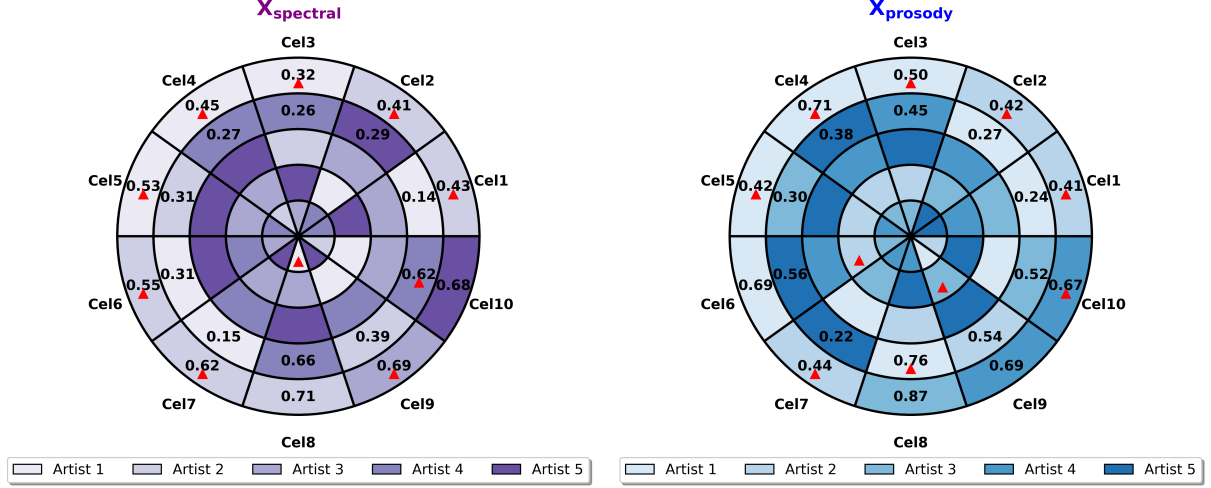


Figure 2: Visualization of similarity scores for ten celebrities (denoted as Cel) across feature sets: X_{spectral} and X_{prosody} . The 360° space is divided into ten wedges, each representing a celebrity. Each wedge is divided into five distinct colored segments (representing five mimicry artists), in the descending order of similarity scores of the artists, with the highest score at the outer segment and the lowest at the center. The best-mimicked artist, based on MOS, is marked with a small red filled triangle ($\blacktriangle$). A color-coded legend for five artists is included. Only ten out of twenty-five celebrities are shown due to space constraints.

3.2. Subjective Evaluation

A subjective assessment with fifteen listeners evaluated the proficiency of mimicry artists. Each listener received recordings of the original celebrity voice alongside mimicked versions by five artists (Artist-1 to Artist-5) and rated their similarity on a 5-point scale (5: very similar, 1: very different). The MOS for each artist is the average of all listener ratings, with the highest MOS indicating the most skilled mimic.

3.3. Experimental Framework

The spectral and prosodic features are extracted from audio files using the librosa toolkit, with a frame size of 30ms and a frame shift of 10ms. The spectral and prosodic features are transformed into X-vector embeddings facilitated by the Speech-Brain Python framework.

An LSTM-based Siamese network is used, with similarity scores serving as an evaluation metric for different artists. The network is trained using a contrastive loss function with a margin of 1.0, and the Adam optimizer is employed with a learning rate of 1×10^{-5} . Training is performed with a batch size of 64 for up to 256 epochs. Ten percentage of data is used for validation. Similarity scores generated by the trained model are used to evaluate the resemblance between original and mimicked voices for each class. The similarity scores obtained by the speaker embeddings X_{spectral} and X_{prosody} are combined using the score fusion technique as follows:

The final score S_f , in the score fusion, is computed as,

$$S_f = \alpha S_{ss} + (1 - \alpha) S_{ps} \quad (1)$$

where S_{ss} , S_{ps} , α represent similarity scores obtained by X-vector speaker embeddings derived from spectral features, prosodic features and weighting constant, respectively. Here alpha (α) will be any value ranging from 0 to 1.

The proposed system ranks artists according to their likelihood of being the most accurate mimic, with Rank 1 assigned to the top candidate. A "hit" occurs when the highest-ranked artist matches the best mimic assessed by humans. Performance is measured using the Top-K hit rate, which tracks how often the

Table 3: The results of Siamese network with MOS. The red-coloured entry is the rank one position. 'Cel' represents Celebrity and 'Art' represents Artist.

Target	Imitator	MOS	Spectral	X_{spectral}	Prosody	X_{prosody}
Cel 1	Art-1	2.77	0.36	0.14	0.36	0.24
	Art-2	3.92	0.31	0.43	0.44	0.41
	Art-3	2.54	0.24	0.07	0.37	0.14
	Art-4	2.84	0.25	0.02	0.40	0.05
	Art-5	1.38	0.14	0.01	0.41	0.01
Cel 2	Art-1	3.62	0.16	0.07	0.32	0.27
	Art-2	3.93	0.31	0.41	0.31	0.42
	Art-3	1.62	0.25	0.26	0.33	0.23
	Art-4	1.54	0.14	0.04	0.29	0.23
	Art-5	1.46	0.80	0.29	0.39	0.21
Cel 3	Art-1	4.38	0.42	0.32	0.24	0.50
	Art-2	3.15	0.04	0.20	0.10	0.40
	Art-3	3.38	0.30	0.04	0.19	0.34
	Art-4	2.46	0.05	0.26	0.27	0.45
	Art-5	2.15	0.01	0.13	0.23	0.41

actual best mimic appears within the Top-K ranks among five artists mimicking twenty-five celebrities. A correct identification within the Top-K earns a point, while an incorrect one receives none.

4. Result Analysis

Due to space limitations, Table 3 shows performance evaluation scores for only three out of twenty-five celebrities. For all three cases, both X_{spectral} and X_{prosody} features based methods correctly identified the best imitator, matching the highest MOS from perceptual evaluations. Specifically, Artist-2 is the best for Celebrities 1 and 2, while Artist-1 is the best for Celebrity 3.

The scores obtained from the feature sets X_{spectral} and X_{prosody} are presented as a radial plot in Figure 2 for ten celebrities. This plot effectively visualizes the highest and second-highest similarity scores of artists imitating a particular celebrity, with annotations marking these values for better clarity. The artist with the best imitation, based on the MOS test, is marked with a small red filled triangle ($\blacktriangle$) in the corresponding section. If the triangle is positioned with the artist having the

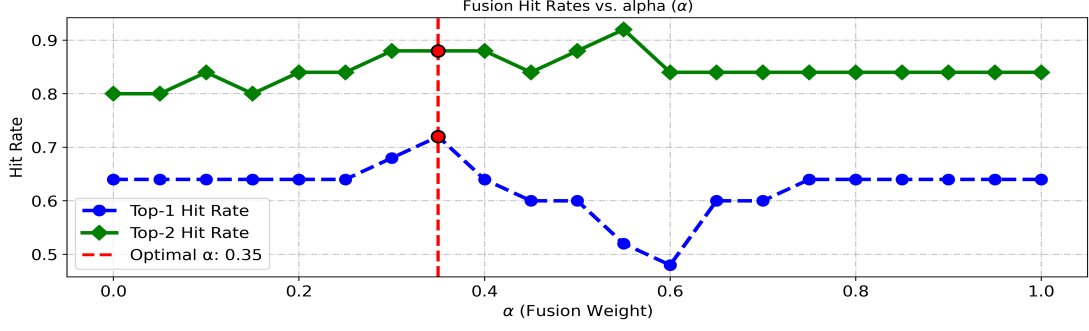


Figure 3: Top-1 and Top-2 hit rates as a function of the fusion weight, (α). The Top-1 and the Top-2 hit rates are plotted for α varying from 0 to 1. The optimal α , where the combined hit rate is maximized, is highlighted with a red dashed vertical line and red markers.

Table 4: Performance comparison of different methods on accuracy in Top-K hit rates (%) on same dataset

No.	Author	Method	Classifier	Loss Function	K=1 (%)	K=2 (%)
1	Abhijith et al. (2020) [4]	$i_{MFCC} + i_{MODGDF} +$ Prosodic (Feature Fusion)	DNN	Categorical cross-entropy	50.00	82.00
2	Bhasi et al. (2024) [25]	Augmented $X_{spectral}$	LSTM (Self Attention)	Categorical cross-entropy	41.00	73.00
3	Proposed Method	$X_{spectral}$	Siamese Network	Binary cross-entropy	40.00	60.00
4	Proposed Method	$X_{prosody}$	Siamese Network	Binary cross-entropy	36.00	60.00
5	Proposed Method	$X_{spectral}$	Siamese Network	Contrastive loss	64.00	84.00
6	Proposed Method	$X_{prosody}$	Siamese Network	Contrastive loss	64.00	80.00

Table 5: Accuracy in Top K-hit rate (%), where ‘.’ represents score fusion. Results of score fusion is highlighted in red color.

Method	K=1	K=2
Spectral features	32.00	56.00
Prosodic features	28.00	56.00
$X_{spectral}$	64.00	84.00
$X_{prosody}$	64.00	80.00
$X_{spectral} \cdot X_{prosody}$	72.00	88.00

highest similarity score, it corresponds to a Top-1 hit, whereas if the triangle corresponds to the second-best artist, it indicates a Top-2 hit. This representation provides valuable insights into the mimicking abilities of a range of artists.

Table 4 compares the performance of various voice mimicry identification methods using Top-K hit rates, highlighting the superiority of the Siamese Network with contrastive loss over baseline methods. It achieves a higher Top-1 hit rate for $X_{spectral}$ and $X_{prosody}$ and an improved Top-2 hit rate for $X_{spectral}$, while binary cross-entropy results in weaker performance, reinforcing the benefits of contrastive loss.

Table 5 summarizes the results, showing that $X_{spectral}$ and $X_{prosody}$ methods identify the Top-1 and Top-2 artists in (16, 21) and (16, 20) cases, respectively. It can be seen that fusion of these features leads to a significant improvement in the Top-K hit rate, surpassing all baseline methods.

Figure 3 illustrates the score fusion results between $X_{spectral}$ and $X_{prosody}$ for varying α values. The Top-1 hit rate peaks at 0.72 for $\alpha = 0.35$, while the Top-2 hit rate reaches 0.92 at $\alpha =$

0.55 before stabilizing at 0.84. The optimal fusion weight is found at $\alpha = 0.35$, which maximizes the combined hit rates, yielding a Top-1 hit rate of 0.72 and a Top-2 hit rate of 0.88, significantly improving overall performance. These results suggest that a balanced fusion of speaker embeddings from spectral and prosodic features improves identification accuracy, while excessive reliance on a single scoring method at $\alpha = 0.60$, degrades performance, highlighting the need for fine-tuning the fusion weight. The experimental analysis shows that the X-vector-based method outperforms other approaches, mainly due to the ability of TDNN to extract speaker discriminative features and capture long-term speech dependencies. The statistical pooling layer further strengthens this advantage by converting variable-length TDNN outputs into fixed-length X-vectors, ensuring consistent speaker embeddings. Additionally, the contrastive loss function enhances speaker separation by minimizing intra-class variance and maximizing inter-class variance, thereby improving identification accuracy and robustness.

5. Conclusion

Our proposed framework effectively evaluates voice mimicry using X-vector speaker embeddings and a Siamese Network for similarity assessment. Strong agreement between perceptual and feature based assessments confirms its reliability. Contrastive loss refines the learned representations, making them more discriminative. Fusion of X-vector embeddings from spectral and prosodic features enhances the Top-K hit rate, with optimal weights maximizing accuracy. These insights underscore the framework’s precision in quantifying voice mimicry.

6. References

- [1] S. Chen, K. Ren, S. Piao, C. Wang, Q. Wang, J. Weng, L. Su, and A. Mohaisen, "You can hear but you cannot steal: Defending against voice impersonation attacks on smartphones," in *proc. of IEEE 37th Int. Conference on Distributed Computing Systems*, pp. 183–195, 2017.
- [2] M. Farrús, M. Wagner, J. Anguita, and J. Hernando, "Robustness of prosodic features to voice imitation," *9th Annual Conference of the Int. Speech Communication Association*, pp. 613–616, 2008.
- [3] C.-L. Tai, H.-S. Lee, Y. Tsao, and H.-M. Wang, "Filter-based discriminative autoencoders for children speech recognition," in *Proc. of European Conference on Signal Processing*, pp. 449–453, 2022.
- [4] G. Abhijith, S. Adharsh, L. Akshay, and R. Rajan, "Competency evaluation in voice mimicking using acoustic cues," in *proceedings of Interspeech 2020*, pp. 1096–1100, 10 2020.
- [5] M. V. Unnikrishnan and R. Rajan, "Mimicking voice recognition using mfcc-gmm framework," *2017 International Conference on Trends in Electronics and Informatics (ICEI)*, pp. 301–304, 2017.
- [6] R. Rajan, A. Noumida, and S. Sreelakshmi, "Paraconsistent feature analysis for the competency evaluation of voice impersonation," in *proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–7, 2023.
- [7] L. Mary, K. Anish Babu, and A. Joseph, "Analysis and detection of mimicked speech based on prosodic features," *International Journal of Speech Technology*, vol. 15, pp. 407–417, 2012.
- [8] V. Vestman, T. Kinnunen, R. G. Hautamäki, and M. Sahidullah, "Voice mimicry attacks assisted by automatic speaker verification," *Computer Speech & Language*, vol. 59, pp. 36–54, 2020.
- [9] C. Zhang and T. Tan, "Voice disguise and automatic speaker recognition," *Forensic Science International*, vol. 175, no. 2, pp. 118–122, 2008.
- [10] R. Rajan, A. Girish, A. Sabu, and A. Prasannan Latha, "Evaluation of voice mimicking using i-vector framework," in *Proceedings of Speech and Computer (Springer International Publishing)*, pp. 446–456, 2020.
- [11] Y. Gao, R. Singh, and B. Raj, "Voice impersonation using generative adversarial networks," in *proceedings of IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP) 2018*, pp. 2506–2510, 2018.
- [12] Y. Zhang, M. Yu, N. Li, C. Yu, J. Cui, and D. Yu, "Seq2seq attentional siamese neural networks for text-dependent speaker verification," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2019*, pp. 6131–6135, 2019.
- [13] A. Qavi, S. A. Khan, and K. Basir, "Voice morphing based on spectral features and prosodic modification," in *proceedings of 17th IEEE Int. Multi Topic Conference*, pp. 401–405, 2014.
- [14] Y. Zhang and L. Liu, "Multi-task learning for x-vector based speaker recognition," *International Journal of Speech Technology*, vol. 26, pp. 817–823, 2023.
- [15] B. M. L. Srivastava, N. Tomashenko, X. Wang, E. Vincent, J. Yamagishi, M. Maouche, A. Bellet, and M. Tommasi, "Design choices for x-vector based speaker anonymization," in *proceedings of Interspeech 2020*, pp. 1713–1717, 2020.
- [16] J. B. Mala, S. M. A. Raj, and R. Rajan, "X-Vector-Based Speaker Diarization Using Bi-LSTM and Interim Voting-Driven Post-processing," in *proceedings of Text, Speech, and Dialogue*, pp. 161–173, 2024.
- [17] B. M. L. Srivastava, M. Maouche, M. Sahidullah, E. Vincent, A. Bellet, M. Tommasi, N. Tomashenko, X. Wang, and J. Yamagishi, "Privacy and utility of x-vector based speaker anonymization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2383–2395, 2022.
- [18] P. Sandhya, V. Spoorthy, S. G. Koolagudi, and N. Sobhana, "Spectral features for emotional speaker recognition," *2020 Third International Conference on Advances in Electronics, Computers and Communications (ICAEECC)*, pp. 1–6, 2020.
- [19] A. Jacob and P. Mythili, "Prosodic feature based speech emotion recognition at segmental and supra segmental levels," *2015 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES)*, pp. 1–5, 2015.
- [20] M. Karafiát *et al.*, "Analysis of x-vectors for low-resource speech recognition," in *proceedings of IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP) 2021*, pp. 6998–7002, 2021.
- [21] M. Rouvier, R. Dufour, and P.-M. Bousquet, "Review of different robust x-vector extractors for speaker verification," in *proceedings of 28th European Signal Processing Conference (EUSIPCO) 2020*, pp. 1–5, 2020.
- [22] Y. Zhang, X. Song, and J. Zhang, "Speaker diarization for multi-speaker conversations via x-vectors," in *proceedings of IEEE Int. Conference on Signal, Information and Data Processing (ICSIDP) 2019*, pp. 1–4, 2019.
- [23] L. Zhou, M. Wang, Y. Qian, H. Luo, H. Li, and X. Lin, "Text-independent speaker recognition based on x-vector," in *proceedings of 7th Int. Conference on Signal and Image Processing (ICSIP) 2022*, pp. 121–125, 2022.
- [24] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *proceedings of IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP) 2018*, pp. 5329–5333, 2018.
- [25] K.C.Bhasi, R. Rajan, and A. Noumida, "Attention-augmented X-vectors for the Evaluation of Mimicked Speech Using Sparse Autoencoder-LSTM framework," in *proceedings of Interspeech 2024*, pp. 3804–3808, 9 2024.
- [26] Z. Li and P. Song, "Audio similarity detection algorithm based on siamese lstm network," *2021 6th International Conference on Intelligent Computing and Signal Processing (ICSP)*, pp. 182–186, 2021.
- [27] K. K. Shenoy, M. M. Pai Manohara, and R. M. Pai, "Siamese network based anomaly detection framework for robust federated learning," *2023 3rd International Conference on Intelligent Technologies (CONIT)*, pp. 1–6, 2023.
- [28] K. C. Bhasi and R. Rajan, "MIMICz-an audio corpus for the competency evaluation of voice mimicking," *2024 Third International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT)*, pp. 1–5, 2024.