A ViT-LSTM-Based Compensation Framework for Robust Speaker Verification under Emotional and Mask-Muffled Conditions

Ziqing Yang and Houwei Cao
Department of Computer Science, New York Institute of Technology
{zyang23@nyit.edu, hcao02@nyit.edu}

***Abstract*—In this study, we systematically investigate the impact of emotional expression and face mask types on the performance of automatic speaker verification (ASV) systems. To address the performance degradation caused by these variabilities, we propose a general compensation framework that separates the speaker identity from emotion- and mask-related distortions in speaker embeddings. Our approach leverages a deep neural network architecture that projects embeddings into distinct variability spaces, using eGeMAPS features for emotion compensation and Vision Transformer (ViT) representations from multi-channel spectrograms for mask compensation. This approach effectively separates speaker-specific characteristics from task-related nuisance factors. We evaluate the proposed compensation models on the CREMA-D dataset, demonstrating their robustness and effectiveness in mitigating the negative effects of emotional and muffled speech. The proposed models can be integrated into both end-to-end and stage-wise ASV systems, enhancing the generalizability of speaker embeddings in challenging conditions.**

***Index Terms*—Speech Verification, Compensation Technique, Face Mask, Spectrogram, ViT**

I. Introduction

The rapid advancement of AI and mobile technology has increased the importance of robust speaker verification. While smart and mobile devices offer greater convenience, they also present challenges for speaker verification systems—especially when there are mismatches between enrollment and test recording conditions. These systems typically rely on speaker embeddings to represent individual speakers. However, in real-world scenarios, it is difficult to ensure that test recordings match enrollment conditions. Variability may arise from a range of factors, including internal influences such as emotional state or health conditions, and external influences such as recording environment, the use of face masks, etc.

In this study, we propose general solutions to verify speakers whose speech is influenced by emotional expression and / or muffled by face masks. A key challenge in such scenarios is the variability mismatch between enrollment and test speech. Enrollment is typically conducted using clear, neutral speech without obstructions. In contrast, test speech may contain emotional cues or acoustic distortions caused by face masks, affecting the speech signal. These discrepancies lead to mismatches with enrollment data, reducing the reliability and effectiveness of conventional speaker verification systems.

Several recent studies have addressed the impact of face masks on automatic speaker verification (ASV). For example, Magee et al. [1] observed that different types of face masks attenuate frequencies above 3–5 kHz, reducing speech intelligibility and negatively affecting ASV performance. Khan et al. [2] proposed a peak norm filter to mitigate the effects of mask-muffled speech under various microphone conditions. Similarly, Chen et al. [3] developed a ResNet-based system with multi-head attention to improve robustness against masked speech. Beyond external factors such as face masks, internal factors—particularly the speaker's emotional state—can also degrade ASV performance [4]. Earlier work addressed this mismatch using feature-level conversion and score normalization techniques. More recent approaches have focused on modeling the relationship between speaker identity and emotional variability through compensation methods [4], [5], [6], [7]. For instance, Campbell et al. [6] introduced a Nuisance Attribute Projection (NAP) method for channel compensation in GMM-SVM-based systems. Sarma et al. [8] later proposed a DNN-based approach to compensate for emotional variability in I-vector representations.

In this study, we systematically investigate the effects of different emotional states and various types of face masks on automatic speaker verification (ASV). To mitigate the variability they introduce, we develop robust ASV systems equipped with dedicated compensation mechanisms. Specifically, we propose an advanced DNN-based framework that projects speaker embeddings into targeted variability subspaces.

For emotion compensation, our objective is to separate speaker identity from emotion-related variability in raw embeddings. To this end, we augment speaker embeddings with the hand-crafted acoustic feature set eGeMAPS, which serves as an explicit projection of the emotion subspace and helps reduce emotion-induced distortion. To compensate for face mask–induced variability, we introduce a ViT–BiLSTM-based model. Unlike the emotion compensation approach, which relies on hand-crafted cues, this model employs a pre-trained Vision Transformer (ViT) to extract mask-related representations from multi-channel linear-scale spectrograms of mask-muffled speech. A BiLSTM layer then captures temporal dependencies and produces a learnable projection of mask-related variability. Our contributions are threefold:

- **Impact Analysis:** We analyze how emotional and mask-

muffled speech affect the performance of various speaker verification systems.
- **Proposed Compensation Methods:** We introduce two novel compensation strategies targeting emotional and face mask-induced variability.
- **Integration and Generalizability:** The proposed compensation models can be integrated into both end-to-end and stage-wise ASV pipelines, improving the generalizability of speaker embeddings across complex real-world conditions.

II. Related Work

- ***Vision Transformer (ViT)*:** Dosovitskiy et al. [9] introduced the Vision Transformer (ViT), which adapts the transformer architecture for image classification. It divides an image into non-overlapping patches using a $p \times p$ convolutional kernel with stride p, where p is the patch size. ViT adds 1D positional embeddings to retain spatial information and includes a class token to summarize global features.
- ***RawNet3*:** RawNet3 is an end-to-end speaker verification model that merges elements from RawNet2 [10] and ECAPA-TDNN [11]. It features a Res2Net-based module and multi-layer aggregation, achieving strong results on VoxCeleb1 [12] and VoxCeleb2 [13]. We adopt the ESPnet-SLU implementation from [14].
- ***TitaNet*:** TitaNet is a scalable speaker verification model with an encoder-decoder structure [15]. It uses 1D depthwise separable convolutions and Squeeze-and-Excitation (SE) layers to capture utterance-level context. In this study, we use the TitaNet-L variant with 25.3 million parameters as a pre-trained embedding model.

III. Method

We propose a comprehensive solution for speaker verification in the presence of emotional variability and mask-muffled speech. The proposed compensation model integrates seamlessly into both end-to-end and stage-wise speaker verification frameworks. Figure 1 illustrates the overall architecture of the model and its integration within these systems.

A. Emotion Compensation Model

Previous studies have shown that speaker-based utterance-level representations, such as i-vectors and x-vectors, can capture both speaking style and emotional content [4]. In speaker verification tasks, the presence of emotion-related variability often degrades system performance. To address this, Sarma et al. [8] proposed a DNN-based method to extract emotion-invariant embeddings from compensated i-vectors, thereby improving performance in speaker identification. Their model employs a three-layer DNN to transform emotionally affected i-vectors into their neutral equivalents.

In this study, we propose a more flexible emotion compensation framework. Instead of relying on i-vectors, we use raw speaker embeddings as input, making the approach compatible with a wide range of embedding extractors and architectures. To disentangle speaker identity from emotional variability, we incorporate the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS)—an 88-dimensional expert-designed acoustic feature set—as a projection of the emotion subspace.

The model is trained by taking embeddings from emotional speech as input and embeddings from corresponding neutral speech as the target. The resulting compensated embeddings can be seamlessly integrated into downstream ASV systems. An overview of the proposed architecture is shown in Figure 1.

B. ViT-based Mask Compensation Model

We introduce a versatile solution for face mask compensation in speaker verification by separating speaker-specific information from mask-induced distortions through a learnable projection of mask-related variability. Our method begins by extracting multi-channel linear-scale spectrograms from mask-muffled speech. Following prior work [16], we generate single-channel spectrograms using window sizes of 8, 15, and 30 milliseconds and stack them along the channel axis, improving time–frequency resolution.

To obtain high-level representations, we use a pre-trained Vision Transformer (ViT) [9], which employs multi-head self-attention to capture internal structure and long-range dependencies among spectrogram patches. Instead of using the pooled output, we take the final hidden layer representation to retain local details relevant to mask variability. This representation is then fed into a bidirectional LSTM (BiLSTM), which models temporal structure and produces a learnable projection of mask-induced variability. The projection is concatenated with the original speaker embedding and passed through a three-layer DNN for final compensation. The overall architecture is shown in Figure 1. In this work, we use the ViT-H/14 model pre-trained on ImageNet-21k [17].

C. Stage-wise Speaker Verification System

In this study, we adopt the conventional X-vector/PLDA pipeline as the baseline model for stage-wise speaker verification. The overall structure of the baseline framework is shown in Figure 1. In the front end, Mel-frequency cepstral coefficients (MFCCs) and voice activity detection (VAD) features are extracted from each audio file. These features are used to adapt a pre-trained time-delay neural network (TDNN) for X-vector extraction following Snyder et al. [18]. Specifically, we use the pre-trained SITW model provided by the Kaldi toolkit [19], trained on the augmented VoxCeleb1 [12] and VoxCeleb2 [13] datasets. The resulting raw speaker embeddings are then processed using our proposed mask compensation model. In the back end, linear discriminant analysis (LDA) [20] and within-class covariance normalization (WCCN) [21] are applied for channel compensation. After length normalization, the compensated X-vectors are used to train a probabilistic linear discriminant analysis (PLDA) model [22] for final speaker verification.

D. End-to-End Speaker Verification System

The proposed mask compensation model can also be seamlessly integrated into end-to-end architectures by inserting

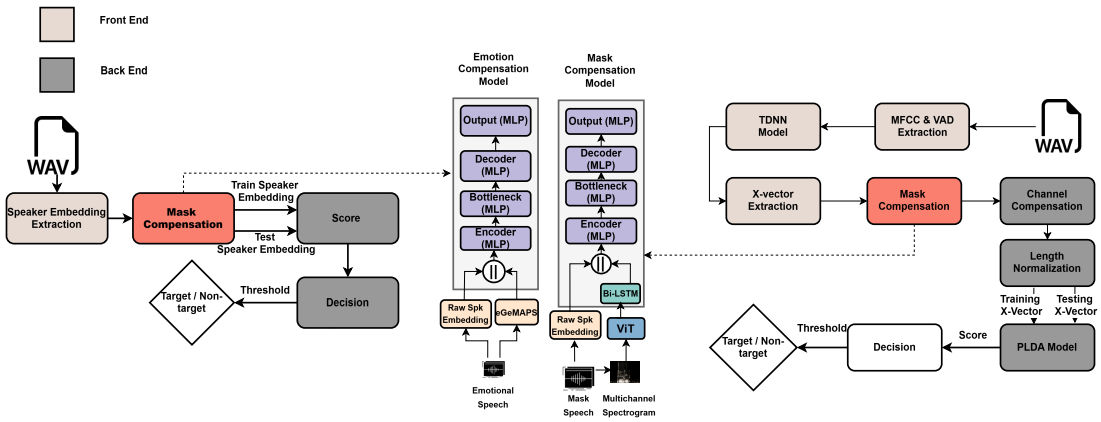


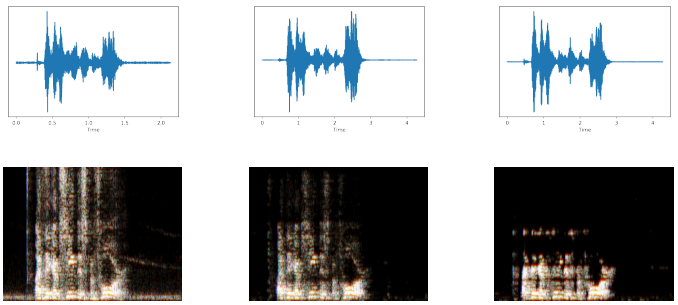
Fig. 1: End-to-End and Stagewise speaker verification system with mask compensation.

the compensation block directly after the speaker embedding layer. As illustrated in Figure 1, the compensated embedding replaces the original one for scoring and decision-making. To evaluate the effectiveness of our method in end-to-end settings, we integrate it with two leading speaker verification systems—RawNet3 [14] and TitaNet [15]—both pre-trained on VoxCeleb1 [12]. These systems achieve state-of-the-art performance, providing a strong testbed for assessing the benefits of our compensation approach.

IV. Experiments & Results

A. Dataset

In this study, we utilize the CREMA-D dataset [23], an audiovisual emotion corpus comprising 7,442 video clips totaling over 10 hours. The clips feature 91 speakers of diverse ethnic backgrounds, with each speaker delivering 12 distinct sentences across six basic emotions: anger, disgust, fear, happiness, neutral, and sadness.



(a) NoMask (b) M_Fabric (c) M_Surgical

Fig. 2: Example speech waveform and Multichannel Spectrogram for original speech without mask (left), re-generated speech with surgical mask (middle), and fabric mask (right).

For our experiments, we focus exclusively on the audio modality. To simulate mask-muffled speech, we recreated audio using two types of face masks—a fabric mask and a disposable surgical mask [24]. Original audio clips were played through a Bose SoundLink Micro Bluetooth speaker covered by each mask, and the resulting muffled signals were recorded with a digital sound recorder. This process produced three versions of every utterance: clean speech (**No_Mask**), speech with a surgical mask (**M_Surgical**), and speech with a fabric mask (**M_Fabric**). Figure 2 shows example waveforms and multichannel spectrograms, highlighting the high-frequency attenuation caused by the masks.

We evaluate the proposed mask-invariant speaker verification models using these datasets. Each speaker contributes 12 utterances (∼36 secs) per emotion. As shown in Table I, training uses 6–11 utterances (∼18–33 secs), while the remaining 1–6 utterances (∼3–18 secs) are reserved for testing.

TABLE I: Speaker models trained and tested with various amount of training and testing data.

Speaker Models	Training Data	Testing Data
33s	11 utterances (∼33 secs)	1 utterance (∼3 secs)
30s	10 utterances (∼30 secs)	2 utterances (∼6 secs)
27s	9 utterances (∼27 secs)	3 utterances (∼9 secs)
24s	8 utterances (∼24 secs)	4 utterances (∼12 secs)
21s	7 utterances (∼21 secs)	5 utterances (∼15 secs)
18s	6 utterances (∼18 secs)	6 utterances (∼18 secs)

B. Within-Emotion Speaker Verification on Clean Speech

Table II summarizes the performance of various speaker verification (SV) systems on clean speech for within-emotion experiments, where the training and testing utterances share the same emotional state. All systems perform better on neutral speech than on emotional speech, indicating that emotional variability introduces additional challenges for SV. We also find that the stage-wise X-Vector/PLDA system is highly sensitive to the amount of training data, as its PLDA backend is trained from scratch and strongly depends on the size and characteristics of the training set. A pre-trained PLDA backend was also examined but did not perform comparably to the version trained from scratch.

C. Cross-Emotion Speaker Verification on Clean Speech

Table III reports cross-emotion results, where models are trained on neutral speech and tested on emotional speech. As expected, all systems show degraded performance compared to

TABLE II: Equal Error Rate (EER%) results on clean speech (**No_Mask**) for multiple speaker verification (SV) systems trained with varying amounts of within-emotion data.

SV System	Amount of Training data					
	33s	30s	27s	24s	21s	18s
EER% on Neutral Speech						
X-Vector/PLDA	4.92	7.02	7.04	10.68	25.47	33.21
RawNet3	3.92	4.31	4.38	4.25	4.02	3.76
TitaNet	5.12	5.14	5.41	5.19	4.78	4.28
EER% on Emotional Speech						
X-Vector/PLDA	5.52	11.96	14.2	19.02	28.87	36.88
RawNet3	8.18	8.80	9.23	9.08	8.80	8.64
TitaNet	8.79	10.39	10.32	10.38	10.01	9.38

the within-emotion results in Table II, reflecting the challenge posed by emotional mismatch.

For the X-vector/PLDA baseline, the performance gap between within- and cross-emotion settings narrows as training data decreases, suggesting that emotion-specific training becomes less impactful in low-resource scenarios and that the model struggles to generalize emotional variability.

We then evaluate the proposed emotion compensation models. Across all systems, emotion-compensated embeddings consistently outperform the cross-emotion baseline. Among them, **RawNet3** shows the greatest improvement. Notably, for both end-to-end systems—RawNet3 and TitaNet—the compensated embeddings even surpass their within-emotion performance, indicating that the compensation model not only mitigates emotional variability but also enhances the overall robustness of speaker representations.

TABLE III: Equal Error Rate (EER%) results for cross-emotion experiments across multiple SV systems trained with varying amounts of clean speech (**No_Mask**)

SV System	Amount of Training data					
	33s	30s	27s	24s	21s	18s
Cross Emotion EER%						
X-Vector/PLDA	14.30	16.29	16.25	21.05	32.34	37.36
RawNet3	11.20	11.70	10.10	9.81	9.50	8.93
TitaNet	11.91	13.01	10.97	11.02	10.13	10.10
Emotion Compensation Embedding EER%						
X-Vector/PLDA	8.71	15.10	13.60	15.16	27.28	33.75
RawNet3	**5.44**	**8.03**	**7.26**	**6.56**	**6.29**	**6.07**
TitaNet	6.55	10.20	8.66	7.53	8.03	6.96

D. Speaker Verification with Mask-muffled Speech

We now examine speaker verification under mask-muffled speech. To systematically evaluate the impact of different face masks, we conduct two groups of experiments: within-corpus and cross-corpus. In the within-corpus setting, both training and testing utterances come from the same corpus (i.e., recorded under the same mask condition). In the cross-corpus setting, models are trained on clean speech but tested on mask-muffled speech. In both cases, the emotional states of the training and testing utterances are always matched.

TABLE IV: Equal Error Rate (EER%) for Fabric **(M_Fabric)** and Surgical mask **(M_Surgical)** conditions across multiple SV systems trained with varying amounts of data.

SV System	Amount of Training data					
	33s	30s	27s	24s	21s	18s
Fabric Face Mask Within Corpus EER%						
X-Vector/PLDA	**17.15**	21.98	24.79	29.33	36.06	42.92
RawNet3	24.85	27.21	26.91	26.98	26.28	25.53
TitaNet	24.21	26.11	25.36	25.84	25.83	25.25
Fabric Face Mask Cross Corpus EER%						
X-Vector/PLDA	35.35	37.42	39.53	41.66	45.44	46.85
RawNet3	32.63	33.10	33.94	34.24	34.07	33.64
TitaNet	34.09	34.64	35.87	35.77	35.17	34.60
Fabric Face Mask Compensation Embedding EER%						
X-Vector/PLDA	17.61	20.44	21.22	24.57	30.64	39.77
RawNet3	19.46	**19.78**	**20.20**	**20.61**	**20.26**	**20.20**
TitaNet	23.13	23.79	24.10	23.74	23.13	23.31
Surgical Face Mask Within Corpus EER%						
X-Vector/PLDA	33.39	36.42	38.85	40.71	44.49	46.40
RawNet3	32.63	35.88	35.56	35.98	35.46	34.38
TitaNet	36.18	36.95	36.91	37.64	38.16	37.46
Surgical Face Mask Cross Corpus EER%						
X-Vector/PLDA	38.61	41.19	41.49	43.57	45.80	47.30
RawNet3	34.16	35.33	36.97	37.22	36.95	36.67
TitaNet	36.73	36.76	38.32	38.34	37.57	37.53
Surgical Face Mask Compensation Embedding EER%						
X-Vector/PLDA	22.17	25.3	29.76	26.08	33.05	39.52
RawNet3	**22.07**	**23.33**	**24.69**	**23.90**	**23.98**	**23.27**
TitaNet	26.74	27.81	28.61	27.94	27.53	27.64

Tables IV present the speaker verification (SV) results for fabric and surgical mask conditions, respectively. Compared with the clean-speech results in Table II, it is evident that face masks significantly degrade SV performance in both within-corpus and cross-corpus settings. All three SV systems show reduced accuracy, with surgical masks causing the most pronounced performance decline.

We first evaluate the effectiveness of the compensation

models on fabric masks. As shown in Table IV, a clear performance gap exists between within-corpus and cross-corpus settings: embeddings extracted from clean speech perform poorly when tested on fabric mask–muffled speech. In contrast, the compensated embeddings consistently outperform the original embeddings from both clean and masked speech. Among the three systems, **RawNet3** performs best in most cases, though with sufficient training data the X-vector/PLDA model becomes comparable or even superior. The pre-trained end-to-end models offer no clear advantage over the stage-wise baseline, highlighting the effectiveness of our compensation approach in improving embedding generalizability.

For surgical masks, performance declines further. As illustrated in Figure 2, surgical masks (**M_Surgical**) introduce stronger high-frequency attenuation, leading to worse SV performance across all systems. The mismatch between clean-speech training and surgical mask testing severely impacts the X-vector/PLDA model. With limited training data, its EER approaches random guessing. Still, the compensated embeddings again outperform the original embeddings derived from either clean or mask-muffled speech. **RawNet3** achieves the best results, although the stage-wise X-vector/PLDA model remains competitive—and can surpass TitaNet—when 24 seconds or more of training data are available.

TABLE V: Cross-emotion speaker verification results (EER%) on fabric mask-muffled speech (**M_Fabric**).

SV System	Amount of Training data					
	33s	30s	27s	24s	21s	18s
Cross-Emotion, Within-Corpus EER%						
RawNet3	29.54	30.92	29.67	28.70	28.01	27.71
TitaNet	28.65	29.70	28.31	28.24	28.10	28.11
Cross-Emotion, Cross-Corpus EER%						
RawNet3	34.20	34.51	33.43	32.94	32.84	32.60
TitaNet	34.38	35.61	34.75	35.06	35.29	35.13
Cross-Emotion, Fabric Face Mask Compensation EER%						
RawNet3	21.61	22.70	20.84	20.02	19.36	19.78
TitaNet	25.18	25.61	23.37	23.51	23.26	24.09
Emotional & Fabric Face Mask Compensation EER%						
RawNet3	**19.28**	**19.61**	**19.11**	**18.44**	**18.21**	**17.99**
TitaNet	21.73	23.39	21.74	21.07	21.56	21.45

E. Cross-Emotion Speaker Verification With Mask-muffled Speech

Finally, we examine the most challenging scenario: cross-emotion speaker verification with mask-muffled speech. Since the end-to-end systems consistently outperform the stage-wise baseline, our analysis of these complex cases focuses only on the two end-to-end speaker verification models.

TABLE VI: Cross-emotion speaker verification results (EER%) on surgical mask-muffled speech (**M_Surgical**).

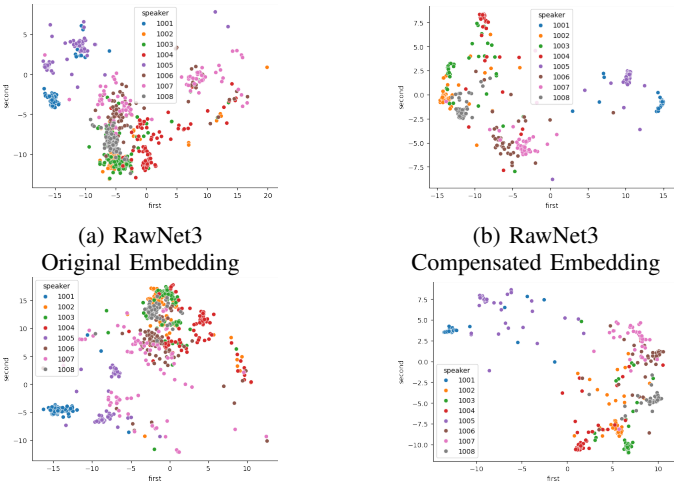
SV System	Amount of Training data					
	33s	30s	27s	24s	21s	18s
Cross-Emotion, Within-Corpus EER%						
RawNet3	37.59	38.63	37.71	37.37	36.93	36.75
TitaNet	39.48	39.59	39.54	38.78	38.61	38.47
Cross-Emotion, Cross Corpus EER%						
RawNet3	35.28	37.83	37.36	37.02	36.95	36.42
TitaNet	38.65	39.94	38.61	38.60	39.02	38.72
Cross-Emotion, Surgical Face Mask Compensation EER%						
RawNet3	24.78	25.06	24.71	23.78	24.16	23.52
TitaNet	29.88	29.67	28.90	28.44	27.75	27.55
Emotional & Surgical Face Mask Compensation EER%						
RawNet3	**22.13**	**23.62**	**23.44**	**21.80**	**21.85**	**22.04**
TitaNet	27.09	29.12	27.78	26.77	26.60	25.94

The results in Tables V and VI highlight the challenges of speaker verification under combined cross-emotion and mask-muffled conditions. Compared with the mask-only results in Table IV, the added emotional variability further degrades performance across all settings. As expected, embeddings derived from speech with surgical masks (**M_Surgical**) perform worse than those with fabric masks (**M_Fabric**), likely due to stronger high-frequency attenuation.

Despite these challenges, all compensation models provide substantial improvements. The proposed joint emotion–mask compensation achieves the lowest EERs across all conditions, demonstrating its effectiveness in mitigating compounded variability. Among the two end-to-end systems, **RawNet3** consistently outperforms TitaNet across all training durations and compensation settings, showing the most significant gains. Although both models benefit from the compensation strategies, the improvements are more pronounced for RawNet3. These results confirm that emotional and mask-induced variabilities can be jointly modeled and that compensation significantly enhances the robustness and generalizability of speaker embeddings in complex, real-world scenarios.

F. Embedding Visualization

To further validate the effectiveness of our proposed solution, we visualized the emotional fabric mask–compensated embeddings extracted from the end-to-end systems using t-SNE [25]. Figure 3 shows the t-SNE projections of both the original and compensated embeddings, with eight randomly selected speakers represented by distinct colors. Comparing Figure 3a with Figure 3b, and Figure 3c with Figure 3d, the compensated embeddings form noticeably more compact and well-separated clusters, indicating improved speaker dis-



(a) RawNet3 Original Embedding

(b) RawNet3 Compensated Embedding

(c) TitaNet Origincal Embedding

(d) TitaNet Compensated Embedding

Fig. 3: Embedding Visualization for 8 Speakers: The left side displays the original embeddings; the right side shows the emotional & fabric mask-compensated embeddings from RawNet3 and TitaNet.

criminability. In contrast, the original embeddings exhibit greater overlap and less distinct boundaries. This visualization provides additional evidence that the proposed compensation framework enhances speaker separability under emotional and mask-muffled conditions.

V. Conclusion

In this study, we investigated how emotional expression and face mask usage affect speaker verification performance and proposed compensation techniques to address these sources of variability. Emotion-related variability was mitigated using eGeMAPS-based projections, while mask-induced distortions were handled with a ViT–LSTM model trained on multichannel spectrograms. Evaluations on the CREMA-D dataset under emotional, mask-muffled, and combined conditions show that our methods improve robustness, yielding 17.4%–40% relative EER reductions across multiple SV frameworks. The proposed compensation strategies enhance the generalizability of speaker embeddings and are effective for both stage-wise and end-to-end systems.

References

[1] Michelle Magee, Courtney Lewis, Gustavo Noffs, Reece, et al., "Effects of face masks on acoustic analysis and speech perception: Implications for peri-pandemic protocols," *The Journal of the Acoustical Society of America*, vol. 148, no. 6, pp. 3562–3568, 2020.

[2] Awais Khan, Ali Javed, Khalid Mahmood Malik, Muhammad Anas Raza, James Ryan, Abdul Khader Jilani Saudagar, and Hafiz Malik, "Toward realigning automatic speaker verification in the era of covid-19," *Sensors*, vol. 22, no. 7, pp. 2638, 2022.

[3] Chaotao Chen, Di Jiang, Jinhua Peng, Rongzhong Lian, Chen Jason Zhang, Qian Xu, Lixin Fan, and Qiang Yang, "A health-friendly speaker verification system supporting mask wearing," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, vol. 35, pp. 16004–16006.

[4] Jennifer Williams and Simon King, "Disentangling style factors from speaker representations.," in *Interspeech*, 2019, pp. 3945–3949.

[5] Alex Solomonoff, William M Campbell, and Ian Boardman, "Advances in channel compensation for svm speaker recognition," in *Proceedings.(ICASSP'05).* IEEE, 2005, vol. 1, pp. I–629.

[6] William M Campbell, Douglas E Sturim, Douglas A Reynolds, and Alex Solomonoff, "Svm based speaker verification using a gmm supervector kernel and nap variability compensation," in *2006 IEEE International conference on acoustics speech and signal processing proceedings*. IEEE, 2006, vol. 1, pp. I–I.

[7] Ali Bou Nassif, Ismail Shahin, Nawel Nemmour, Noor Hindawi, and Ashraf Elnagar, "Emotional speaker verification using novel modified capsule neural network," *Mathematics*, vol. 11, no. 2, pp. 459, 2023.

[8] Biswajit Dev Sarma and Rohan Kumar Das, "Emotion invariant speaker embeddings for speaker identification with emotional speech," in *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2020, pp. 610–615.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[10] Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, and Anthony Larcher, "End-to-end anti-spoofing with rawnet2," in *ICASSP 2021-2021*. IEEE, 2021, pp. 6369–6373.

[11] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," in *Proc. Interspeech 2020*, 2020, pp. 3830–3834.

[12] A Nagrani, J Chung, and A Zisserman, "Voxceleb: a large-scale speaker identification dataset," *Interspeech 2017*, 2017.

[13] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "Voxceleb2: Deep speaker recognition," in *Proc. Interspeech 2018*, 2018, pp. 1086–1090.

[14] Jee-weon Jung, Wangyou Zhang, Jiatong Shi, Zakaria Aldeneh, Takuya Higuchi, Alex Gichamba, Barry-John Theobald, Ahmed Hussen Abdelaziz, and Shinji Watanabe, "Espnet-spk: full pipeline speaker embedding toolkit with reproducible recipes, self-supervised front-ends, and off-the-shelf models," in *Proc. Interspeech 2024*, 2024, pp. 4278–4282.

[15] Nithin Rao Koluguri, Taejin Park, and Boris Ginsburg, "Titanet: Neural model for speaker representation with 1d depth-wise separable convolutions and global context," in *ICASSP 2022-2022*. IEEE, 2022, pp. 8102–8106.

[16] Jeno Szep and Salim Hariri, "Paralinguistic classification of mask wearing by image classifiers and fusion.," in *Interspeech*, 2020, pp. 2087–2091.

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[18] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 ICASSP*. IEEE, 2018, pp. 5329–5333.

[19] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Burget, et al., "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011, number CONF.

[20] Peter E Hart, David G Stork, and Richard O Duda, *Pattern classification*, Wiley Hoboken, 2000.

[21] Andrew O Hatch, Sachin Kajarekar, and Andreas Stolcke, "Within-class covariance normalization for svm-based speaker recognition," in *Ninth international conference on spoken language processing*, 2006.

[22] Simon JD Prince and James H Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *2007 IEEE 11th international conference on computer vision*. IEEE, 2007, pp. 1–8.

[23] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.

[24] Ziqing Yang, Katherine Nayan, Zehao Fan, and Houwei Cao, "Multimodal emotion recognition with surgical and fabric masks," in *ICASSP 2022-2022*. IEEE, 2022, pp. 4678–4682.

[25] Laurens Van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.