END-TO-END SPEAKER VERIFICATION WITH UNCERTAINTY-AWARE EVIDENTIAL SCORING

Min Hyun Han, Chanyeong Moon, Ju Yeon Kang, and Nam Soo Kim

Department of Electrical and Computer Engineering and INMC
Seoul National University, Seoul, South Korea
{mhhan, cymoon, jykang}@hi.snu.ac.kr, nkim@snu.ac.kr

ABSTRACT

End-to-end speaker verification has emerged as a promising approach for simplifying traditional multi-stage pipelines in speaker verification tasks, enabling direct optimization of speaker embedding and scoring functions. However, conventional methods lack ability to quantify uncertainty, which is crucial in high-stakes applications where reliability and interpretability are essential. In this work, we propose an evidential deep learning-based end-to-end speaker verification model that integrates speaker representation learning, verification scoring, and uncertainty estimation within a single framework. By leveraging evidential deep learning, our model not only improves speaker verification performance but also provides uncertainty estimates. Experimental results on benchmark datasets demonstrate that our approach achieves competitive verification performance while offering uncertainty quantification, enhancing its applicability in real-world scenarios.

***Index Terms*—** Speaker verification, end-to-end scoring, evidential deep learning, uncertainty quantification

1. INTRODUCTION

Speaker verification aims to determine whether two audio samples belong to the same speaker. Traditional systems typically follow a multi-stage pipeline that separates speaker embedding extraction from the scoring process [1]. With the advent of deep learning, the field has undergone significant advancements, particularly in learning speaker-discriminative representations directly from audio. This shift has led to the development of deep speaker embedding frameworks, which leverage deep neural networks to extract fixed-dimensional embeddings that capture speaker characteristics. These embeddings have proven to be effective for speaker verification tasks, offering more accurate and robust performance under diverse conditions [2].

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2025-00554289)

Deep speaker embedding methods have primarily evolved along two paradigms. Classification-based approaches [3–6] optimize embeddings through softmax or margin-based objectives, which provide stable training but often exhibit limited generalization to unseen speakers. Conversely, metric learning methods [7–10] explicitly minimize intra-speaker variability and maximize inter-speaker separability, thereby aligning more directly with verification goals. Nonetheless, both approaches typically require additional back-end scoring modules such as probabilistic linear discriminant analysis (PLDA) [11] or score normalization [12], indicating the continued need for unified and robust end-to-end frameworks.

End-to-end verification strategies have emerged as a robust alternative to traditional multi-stage pipelines in speaker verification. These methods streamline the process by integrating speaker embedding extraction and scoring into a unified framework, eliminating the need for separate back-end classifiers or scoring functions. Several end-to-end methods, such as neural PLDA [13, 14], pseudo-distance learning [15], and graph-based scoring [16], have been proposed to unify embedding extraction and verification. Despite these advancements, most end-to-end systems do not explicitly model predictive uncertainty, which is essential in applications requiring high reliability and interpretability. Conventional softmax-based models tend to be overconfident, as small differences in logits are exaggerated by the exponential function, often resulting in high-confidence predictions even for uncertain inputs. This issue is exacerbated by the maximum likelihood estimation (MLE) objectives such as cross-entropy loss, which focus solely on maximizing target class probabilities without accounting for uncertainty.

To address this issue, we propose an end-to-end speaker verification framework based on evidential deep learning (EDL) [17]. EDL allows the model to estimate both similarity scores and their associated uncertainty by modeling evidential distributions without the need for sampling. This approach enables the system to produce calibrated, uncertainty-aware predictions and to identify unreliable trials. In addition, we incorporate a contrastive loss [18] tailored to speaker verification, encouraging the model to assign higher scores to

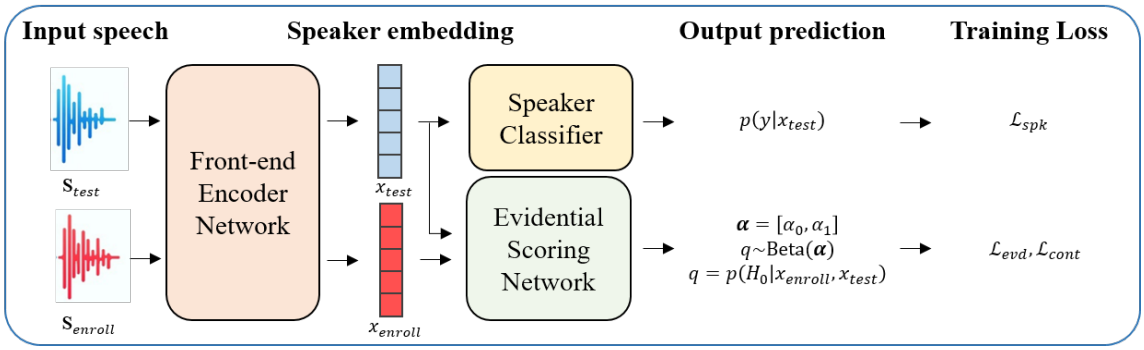

Fig. 1. Overview of the proposed end-to-end deep evidential learning framework for speaker verification. The framework integrates a front-end encoder network with an evidential scoring network, which models both similarity and predictive uncertainty from embedding pairs.

same-speaker pairs than to different-speaker pairs. This further enhances the discriminative power of learned scoring network. Through extensive experiments, we demonstrate that the proposed framework not only improves the verification performance but also provides a comprehensive measure of uncertainty, ensuring that the predictions are both accurate and reliable.

2. PROPOSED METHOD

2.1. Motivation

End-to-end speaker verification simplifies traditional pipelines by jointly learning embedding extraction and scoring, but it lacks mechanisms to quantify predictive uncertainty. In practice, softmax-based models are prone to overconfidence, as small logit differences are exaggerated and cross-entropy training ignores uncertainty. To overcome this, we incorporate evidential deep learning (EDL) [17], which represents outputs as evidential distributions (e.g., Beta or Dirichlet). This enables calibrated, uncertainty-aware decisions without costly sampling, thereby improving robustness and enhancing interpretability for real-world applications.

2.2. Front-end Encoder Network and Speaker Classifier

The front-end encoder network processes the input speech signal and extracts compact speaker embeddings. In our implementation, we employ the ECAPA-TDNN architecture [19] due to its effectiveness in learning speaker-discriminative representations. Given an input speech signal $\mathbf{S}_n$, the front-end encoder network converts it into an utterance-level embedding x_n. During training, the embeddings are optimized through a speaker classifier, which assigns them to their respective speaker classes. The speaker classification loss $\mathcal{L}_{spk}$ is formulated as follows:

$$\mathcal{L}_{spk} = -\frac{1}{N}\sum_{n=1}^{N}\log\frac{\exp(f(x_n, w_{y_n}))}{\sum_{j=1}^{C}\exp(f(x_n, w_j))}, \quad (1)$$

where N is the total number of training utterances, C is the number speakers in the training set, y_n is the speaker label and $f(x_n, w_j)$ represents the similarity function (e.g., cosine similarity or AM-softmax) between the embedding x_n and the j-th speaker basis in the classifier w_j.

2.3. Evidential Scoring Network

To model uncertainty in verification decisions, we introduce the Evidential Scoring Network (ESN), which acts as a binary classifier that distinguishes between the target (same-speaker) hypothesis H_0 and impostor (different-speaker) hypothesis H_1. Given a pair of embeddings, one from the enrollment utterance x_{enr} and the other from the test utterance x_{tst}, the ESN outputs two non-negative parameters $\boldsymbol{\alpha} = [\alpha_0, \alpha_1]$ $(\alpha_k \geq 1)$, which represent evidence in favor of each hypothesis. These parameters define a Beta distribution over the verification probability $q = p(H_0|x_{tst}, x_{enr})$:

$$Beta(q; \alpha_0, \alpha_1) = \frac{1}{B(\boldsymbol{\alpha})}q^{\alpha_0 - 1}(1-q)^{\alpha_1 - 1} \quad (2)$$

where $B(\boldsymbol{\alpha})$ is the Beta function that normalizes the distribution. This formulation allows the model to represent not just a point estimate but a full distribution over the verification probability, which quantifies both the predicted probability and its associated uncertainty. During the inference phase, the expected probability of this distribution is used as the verification score:

$$\hat{p}(H_0|x_{tst}, x_{enr}) = E[q] = \frac{\alpha_0}{\alpha_0 + \alpha_1}. \quad (3)$$

The associated predictive uncertainty is calculated as:

$$u = \frac{2}{\alpha_0 + \alpha_1}, \quad (4)$$

which is inversely proportional to the total amount of evidence. When the model produces a little evidence, the uncertainty is considered high. Conversely, strong and confident predictions yield higher evidence and lower uncertainty.

2.4. Training Loss Functions

The evidential scoring network (ESN) is trained to estimate both similarity and predictive uncertainty between speaker embeddings. Let a training batch consist of N speakers, each with two utterances. For speaker i, the first utterance is denoted as $x_{\text{tst}}^{(i)}$ and the second as $x_{\text{enr}}^{(j)}$. We construct N^2 test-enroll pairs $(x_{\text{tst}}^{(i)}, x_{\text{enr}}^{(j)})$ each associated with a label $l_{ij} \in \{0, 1\}$, indicating whether the pair comes from the same speaker or not.

The ESN outputs Beta distribution parameters $\boldsymbol{\alpha}_{ij} = [\alpha_0^{(ij)}, \alpha_1^{(ij))}]$, from which the predictive distribution over the verification probability $q = p(H_0|x_{tst}^{(i)}, x_{enr}^{(j)})$ is defined as:

$$Beta(q; \alpha_0^{(ij)}, \alpha_1^{(ij)}) = \frac{1}{B(\boldsymbol{\alpha}_{ij})} q^{\alpha_0^{(ij)}-1}(1-q)^{\alpha_1^{(ij)}-1}. \quad (5)$$

Instead of minimizing the squared error between a point prediction and the label, we minimize the expected squared error over the predicted Beta distribution, i.e., the Bayes risk

$$\mathcal{L}_{ij} = \mathbb{E}_{q\sim Beta(\boldsymbol{\alpha_{ij}})}[(l_{ij} - q)^2]. \quad (6)$$

This can be analytically computed as follows:

$$\mathcal{L}_{ij} = (l_{ij} - \hat{p}_{ij})^2 + \frac{\hat{p}_{ij}(1-\hat{p}_{ij})}{S_{ij}+1} \quad (7)$$

where $\hat{p}_{ij} = \frac{\alpha_0^{(ij)}}{\alpha_0^{(ij)}+\alpha_1^{(ij)}}$ is the expected probability of the same-speaker hypothesis, and $S_{ij} = \alpha_0^{(ij)} + \alpha_1^{(ij)}$ denotes the Dirichlet strength (i.e., total evidence). The second term acts as a regularizer, penalizing predictions made with low evidence. Aggregating over all N^2 training pairs, the full evidential loss becomes

$$\mathcal{L}_{\text{evd}} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left((l_{ij} - \hat{p}_{ij})^2 + \frac{\hat{p}_{ij}(1-\hat{p}_{ij})}{S_{ij}+1}\right). \quad (8)$$

To enhance the discriminability of the learned scores, we additionally apply a contrastive loss that encourages higher scores for genuine pairs relative to impostor pairs as given by

$$\mathcal{L}_{\text{cont}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(s\cdot\hat{p}_{ii})}{\sum_{j=1}^{N}\exp(s\cdot\hat{p}_{ij})} \quad (9)$$

where s is a scaling factor applied to the scores. The contrastive loss enhances the separability of target and non-target score distributions, leading to more reliable verification decisions.

The total training objective is a weighted sum of the speaker classification loss $\mathcal{L}_{\text{spk}}$, the evidential loss $\mathcal{L}_{\text{evd}}$, and the contrastive loss $\mathcal{L}_{\text{cont}}$:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{spk}} + \lambda_{\text{evd}}\mathcal{L}_{\text{evd}} + \lambda_{\text{cont}}\mathcal{L}_{\text{cont}}. \quad (10)$$

3. EXPERIMENTS

3.1. Datasets

For training, we used the development set of VoxCeleb 1&2 datasets [20, 21], which includes 1,092,009 recordings from 5,994 different speakers. VoxCeleb is widely recognized for large-scale text-independent speaker verification tasks. Speech samples are sourced from YouTube videos that capture real-world scenarios with various types of ambient noise, such as background chatter, laughter, and overlapping speech.

For evaluation, we employed the Vox1-O protocol, which consists of 37,611 trials with 4,708 utterances. In addition, the CN-Celeb (E) evaluation set [22] was used to assess cross-lingual and multi-genre robustness, which contains 18,849 utterances from 200 speakers and involving 3,484,292 trials.

3.2. Network settings

We employed two types of input features: (1) 80-dimensional log Mel-filterbank energies extracted from 2-second speech segments with standard augmentations (MUSAN noise [23], simulated RIRs [24], and speed perturbation), and (2) WavLM-large representations [25] obtained as a learnable weighted combination of all transformer layers. For the encoder, we adopted ECAPA-TDNN [19] (C=1024), and the evidential scoring network (ESN) was implemented as a two-layer fully connected network. Training used the Adam optimizer with cosine-annealed learning rate scheduling and warm-up restarts [26].

3.3. Baseline Back-End Scoring Methods

We compared the proposed evidential scoring method with the following baseline back-end scoring approaches:

- **PLDA** [11]: Generative scoring model that computes log-likelihood ratios.
- **PLDA-diag** [27]: Simplifed PLDA with diagonal covariance for stability and efficiency.
- **ASN** [12]: Score normalization with cohort statistics.
- **NPLDA** [14]: Neural network-based approximation of PLDA scoring.
- **LRD** [15]: Neural network-based scoring model that learns pseudo-distance function.

3.4. Speaker Verification Results

Table 2 presents the speaker verification performance on Vox1-O and CN-Celeb (E). When using the ECAPA-TDNN

Table 1. Performance Metrics by Uncertainty Group

Group	1	2	3	4	5	6	7	8	9	10
uncertainty (u)	0.078∼0.174	0.174∼0.206	0.206∼0.226	0.226∼0.238	0.238∼0.247	0.247∼0.255	0.255∼0.263	0.263∼0.272	0.272∼0.283	0.283∼0.432
EER (%)	11.62	8.96	10.61	10.96	11.50	12.56	12.70	13.18	14.39	25.49
# of target trials	1781	2633	4160	2391	1490	1130	947	884	1004	1335
# of imposter trials	346648	345796	344269	346038	346939	347299	347482	347545	347425	347094

Table 2. Performance comparison across systems on Vox1-O and CN-Celeb (E).

Input	Front-end encoder	Scoring backend	Backend training	Vox1-O		CN-Celeb (E)	
				EER (%)	minDCF	EER (%)	minDCF
Mel-spec.	ECAPA-TDNN	COS	N/A	0.86	0.069	14.27	0.497
Mel-spec.	ECAPA-TDNN	ASN	N/A	0.94	0.086	13.90	0.519
Mel-spec.	ECAPA-TDNN	PLDA	Separate	2.01	0.119	17.71	0.619
Mel-spec.	ECAPA-TDNN	PLDA-diag	Separate	0.98	0.073	15.11	0.543
Mel-spec.	ECAPA-TDNN	LRD	Joint	0.84	0.065	14.27	0.495
Mel-spec.	ECAPA-TDNN	NPLDA	Joint	0.80	0.062	14.02	**0.494**
Mel-spec.	ECAPA-TDNN	ESN (proposed)	Joint	**0.74**	**0.051**	**13.75**	0.497
WavLM	ECAPA-TDNN	COS	N/A	0.65	0.048	21.55	0.591
WavLM	ECAPA-TDNN	ASN	N/A	0.62	0.047	18.95	**0.470**
WavLM	ECAPA-TDNN	ESN (proposed)	Joint	**0.58**	**0.035**	**17.61**	0.512

Table 3. Ablation study on training losses

Loss	Backend	EER	minDCF
$proposed$	COS	0.76	0.0599
	ESN	0.74	0.0510
w/o $\mathcal{L}_{cont}$	COS	0.82	0.0611
	ESN	1.22	0.1010
w/o $\mathcal{L}_{spk}$	COS	1.91	0.1257
	ESN	1.13	0.0895
w/o $\mathcal{L}_{spk}$ & $\mathcal{L}_{cont}$	COS	3.34	0.2745
	ESN	2.83	0.2629
w/o $\mathcal{L}_{evd}$ & $\mathcal{L}_{cont}$	COS	0.86	0.0694

encoder with cosine similarity (COS) as the back-end, the baseline system achieves competitive results. However, integrating the evidential scoring network (ESN) leads to consistent improvements across both datasets. For instance, on Vox1-O, the EER improves from 0.86% to 0.74%, and on CN-Celeb (E), from 14.27% to 13.75%, with similar trends observed in minDCF. These results confirm that the proposed uncertainty-aware scoring method enhances reliability.

We also evaluated the models with WavLM-based features. While WavLM embeddings alone provide strong performance with the COS back-end, integrating ESN still yields noticeable improvements. On CN-Celeb (E), for example, the EER drops from 21.55% to 17.61%, demonstrating that the proposed uncertainty-aware scoring framework generalizes well to stronger feature representations.

Table 3 shows an ablation study of the training loss components. Removing either the contrastive loss ($\mathcal{L}_{cont}$) or the speaker classification loss ($\mathcal{L}_{spk}$) leads to a clear degradation in both EER and minDCF. Using only the evidential loss ($\mathcal{L}_{evd}$) results in the worst performance, confirming that evidential regression without supervision leads to unstable decision boundaries and degraded discriminability. The best results are obtained when all three losses are combined, suggesting that the classification objective stabilizes embedding learning, while the contrastive loss enhances discriminability in the scoring space. These findings highlight the importance of joint optimization in the proposed framework.

3.5. Predictive Uncertainty

Table 1 presents the performance grouped by predictive uncertainty on the CN-Celeb (E) test set. Samples are divided into ten bins based on uncertainty score u, with each group showing the corresponding EER, number of target trials, and imposter trials.

A general trend is observed where higher uncertainty leads to higher EER. Group 1, with the lowest uncertainty, shows an EER of 11.62%, while Group 10, with the highest uncertainty, reaches 25.49%. This supports the notion that uncertainty correlates with prediction reliability.

Notably, Group 2 achieves the lowest EER (8.96%), while Group 1 performs worse despite lower uncertainty. This indicates that some low-uncertainty predictions may still produce overconfident errors.

These results highlight the usefulness of uncertainty estimates in identifying unreliable trials and improving score calibration.

4. CONCLUSION

In this paper, we introduced an end-to-end speaker verification system incorporating deep evidential learning to quantify predictive uncertainty. The integration of the ESN with the ECAPA-TDNN model demonstrated significant performance improvements, reducing error rates and enhancing reliability across multiple test sets. Our ablation study highlighted the effectiveness of combining evidential, contrastive, and speaker classification losses. Additionally, our analysis showed that higher uncertainty generally correlates with poorer performance, while low uncertainty can still lead to overconfidence. These results demonstrate the effectiveness of integrating evidential deep learning for speaker verification, particularly in enhancing robustness and reliability in real-world scenarios.

5. REFERENCES

[1] J. H. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal processing magazine*, vol. 32, no. 6, pp. 74–99, 2015.

[2] Z. Bai and X.-L. Zhang, "Speaker recognition based on deep learning: An overview," *Neural Networks*, vol. 140, pp. 65–99, 2021.

[3] Y. Liu, W.and Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proc. CVPR*, 2017, pp. 212–220.

[4] J. Wang, F. Zhou, S. Wen, X. Liu, and Y. Lin, "Deep metric learning with angular loss," in *Proc. ICCV*. IEEE, 2017.

[5] Z. Huang, S. Wang, and K. Yu, "Angular softmax for short-duration text-independent speaker verification." in *Proc. Interspeech*, 2018, pp. 3623–3627.

[6] Y. Liu, L. He, and J. Liu, "Large margin softmax loss for speaker verification," in *Proc. Interspeech*, 2019, pp. 2873–2877.

[7] C. Zhang, K. Koishida, and J. H. Hansen, "Text-independent speaker verification based on triplet convolutional neural network embeddings," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1633–1644, 2018.

[8] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *Proc. ICASSP*. IEEE, 2018, pp. 4879–4883.

[9] J. Wang, K.-C. Wang, M. Law, F. Rudzicz, and M. Brudno, "Centroid-based deep metric learning for speaker recognition," in *Proc. ICASSP*. IEEE, 2019, pp. 3652–3656.

[10] J. S. Chung, J. Huh, S. Mun, M. Lee, H.-S. Heo, S. Choe, C. Ham, S. Jung, B.-J. Lee, and I. Han, "In defence of metric learning for speaker recognition," in *Proc. Interspeech*, 2020, pp. 2977–2981.

[11] S. Ioffe, "Probabilistic linear discriminant analysis," in *European Conference on Computer Vision*. Springer, 2006, pp. 531–542.

[12] P. Matejka, O. Novotnỳ, O. Plchot, L. Burget, M. D. Sánchez, and J. Cernockỳ, "Analysis of score normalization in multilingual speaker recognition." in *Proc. Interspeech*, 2017, pp. 1567–1571.

[13] S. Ramoji, P. Krishnan, and S. Ganapathy, "Nplda: A deep neural plda model for speaker verification," in *Proc. Odyssey: The Speaker and Language Recongition Workshop*, 2020, pp. 202–209.

[14] ——, "Neural plda modeling for end-to-end speaker verification," in *Proc. Interspeech*, 2020, pp. 4333–4337.

[15] J. Monteiro, I. Albuquerque, J. Alam, R. D. Hjelm, and T. Falk, "An end-to-end approach for the verification problem: learning the right distance," in *Proc. ICML*, 2020, pp. 7022–7033.

[16] J.-w. Jung, H.-S. Heo, H.-J. Yu, and J. S. Chung, "Graph attention networks for speaker verification," in *Proc. ICASSP*. IEEE, 2021, pp. 6149–6153.

[17] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," in *Advances in neural information processing systems*, 2018, pp. 3179–3189.

[18] Y. Tang, J. Wang, X. Qu, and J. Xiao, "Contrastive learning for improving end-to-end speaker verification," in *IJCNN*, 2021, pp. 1–7.

[19] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," in *Proc. Interspeech*, 2020, pp. 3830–3834.

[20] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *Proc. Interspeech*, 2017, pp. 2616–2620.

[21] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Proc. Interspeech*, 2018, pp. 1086–1090.

[22] Y. Fan, J. Kang, L. Li, K. Li, H. Chen, S. Cheng, P. Zhang, Z. Zhou, Y. Cai, and D. Wang, "Cn-celeb: a challenging chinese speaker recognition dataset," in *Proc. ICASSP*. IEEE, 2020, pp. 7604–7608.

[23] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[24] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proc. ICASSP*. IEEE, 2017, pp. 5220–5224.

[25] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[26] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," in *Proc. ICLR*, 2017, pp. 1–13.

[27] Q. Wang, K. A. Lee, and T. Liu, "Scoring of large-margin embeddings for speaker verification: Cosine or plda?" in *Proc. Interspeech*, 2022, pp. 600–604.