FUSION OF EMBEDDINGS NETWORKS FOR ROBUST COMBINATION OF TEXT DEPENDENT AND INDEPENDENT SPEAKER RECOGNITION

Ruirui Li *Chelsea J.-T. Ju* *Zeya Chen* *Hongda Mao* *Oguz Elibol* *Andreas Stolcke*

Amazon Alexa Speech
{ruirul, juitij, zeyachen, hongdam, oelibol, stolcke}@amazon.com

ABSTRACT

By implicitly recognizing a user based on his/her speech input, speaker identification enables many downstream applications, such as personalized system behavior and expedited shopping checkouts. Based on whether the speech content is constrained or not, both text-dependent (TD) and text-independent (TI) speaker recognition models may be used. We wish to combine the advantages of both types of models through an ensemble system to make more reliable predictions. However, any such combined approach has to be robust to incomplete inputs, i.e., when either TD or TI input is missing. As a solution we propose a fusion of embeddings network (FOEnet) architecture, emphasizing joint learning with neural attention. We compare FOEnet with two state-of-the-art baseline methods on a dataset of voice assistant inputs, and show that it achieves higher accuracy than the baseline methods, especially in the presence of incomplete inputs.

1. INTRODUCTION

Speaker recognition answers the question "who is speaking" given a speech utterance. Based on whether the utterance content is constrained or not, there are two types of speaker recognition model: text-dependent (TD) and text-independent (TI). For the former, the utterance content is known or limited to a small set of options. For deployment in voice assistants, the input usually consists of the predefined wakeword, such as "Hey Google", "Echo" or "Alexa". TI systems, on the other hand, make no assumptions on what was said, and therefore have wider applicability. However, they are typically less accurate than TD systems since the speaker's choice of speech content introduces nuisance variability into the signal.

For example, given an utterance like "Alexa, what time is it?", the TD and TI speaker recognition models can predict speaker identity independently. The TD model compares the acoustic signal corresponding to "Alexa" and a user's corresponding enrolled wakeword segments to calculate a matching score. The TI model, on the other hand, can take either the entire utterance or the non-wakeword portion and compare it with users' enrolled utterances to yield another matching score. The two models are generally trained with independent machine learning models and on different datasets with the goal of optimizing their recognition performance individually. When both TD and TI models are available, we should be able to improve overall accuracy by combining their predictions; however, the following three concerns arise. First, it is not clear how to best combine the models, given that a range of fusion methods have been proposed in machine learning (such as early fusion and late fusion [1, 2], etc.). Second, the inputs for the two systems, especially the TD system, may not be always available. For example, the wakeword may be optional in certain contexts. Table 1 further summarizes the scenarios with missing inputs from either the TD or the TI system. Third, a user's voice may change (e.g., due to aging) and they may re-enroll, which may lead to a domain shift between training and test data, which in turn may affect the balance between TI and TD models for fusion purposes.

Speaker recognition has been studied extensively as a supervised classification problem. Traditionally, solutions are based on i-vector representation of speech segments [3]. More recently, methods based on deep neural net embeddings (d-vectors) have been adopted. Deep speaker [4] and VGGVox [5] utilize CNN-based residual networks to extract audio embeddings and optimize triplet and contrastive losses to train speaker recognition models, respectively. Two other Resnet-based methods [6, 7] adopt additive margin softmax classification loss [8] to improve the recognition accuracy. More recently, Mockingjay [9] and TERA [10] learn speaker embedding representations based on self-supervised spectrogram reconstructions. While combining of speaker recognition models was very popular in research systems before d-vectors [11, 12], little recent work has looked at combining multiple neural embedding systems. [13] explores person verification by fusing audio and visual systems. Researchers investigate how to fuse GMM-UBM based TD and TI speaker recognition models in [14], where the TD and TI scores are fed as inputs to a fusion classifier. Nevertheless, it does not address the problem of incomplete inputs.

Our work assumes users have registered profiles consisting of embeddings, each based on four to ten utterances. We also assume pre-trained TD and TI models, which are capable of generating corresponding fix-length embeddings given a new utterance. Unlike the aforementioned fusion research, which combines *score-level* outputs from component systems, we develop a fusion approach that uses *embedding-*

Table 1: Scenarios when speaker profile or utterance embeddings are unavailable. E_{spk} and E_u represent speaker profile and test utterance embedding, respectively. Detailed descriptions of E_{spk}^{TD} and E_u^{TD},E_{spk}^{TI} and E_u^{TI} are given in Section 2.

Scenario	E_{spk}^{TD}	E_u^{TD}	E_{spk}^{TI}	E_u^{TI}
There are no wakewords in an utterance	✓	x	✓	✓
The wakeword detector fails to identify the wakeword in an utterance	✓	x	✓	✓
Devices are woken up by a button press	✓	x	✓	✓
The speaker gets enrolled, but for a different wakeword	x	✓	✓	✓
TD model is working improperly during run time	✓	x	✓	✓
TI model is working improperly during run time	✓	✓	✓	x

level predictions from both TD and TI models. Moreover, we allow the system to operate in the face of incomplete input, such as when an utterance does not start with a wakeword; the wakeword detector fails to identify the wakeword; a user has not yet enrolled for a wakeword sample (but the test utterance contains one); or an unexpected run-time error prevents the recognition engines from generating either TD or TI embeddings.

2. FUSION OF EMBEDDINGS NETWORK

Given a test utterance u, the goal of the speaker recognition model is to learn a scoring function such that the score $s(spk, u)$ between the target speaker and u is as high as possible, while the scores between other speakers and u are as low as possible.

We use E_u^{TD} and E_u^{TI} to denote the embeddings of a test utterance u generated by TD and TI models, respectively. Similarly, we use E_{spk}^{TD} and E_{spk}^{TI} to denote the profile embeddings of speaker spk generated by the TD and TI model, respectively. More precisely, if the profile of a speaker is composed of a set of M utterances $\{u_1, u_2, ..., u_M\}$, then $E_{spk} = \frac{1}{M}\sum_{i=1}^{M} E_{u_i}$.

To effectively fuse predictions from both TD and TI models, while handling the issue of incomplete inputs, we propose the fusion model shown in Fig. 1, which combines TD and TI predictions at the embedding level. When the input from one of the systems is not available, the model will *infer* the missing embedding from the one that is available. This results in a consistent input to the fusion layer, which computes a scalar score from the concatenation of TD and TI embeddings-level outputs.

For both TD and TI models, we first calculate a differential embedding between the speaker profile embedding E_{spk} and the utterance embedding E_u. Formally,

$$E_{diff}^{TD} = E_{spk}^{TD} - E_u^{TD}, \tag{1}$$

and

$$E_{diff}^{TI} = E_{spk}^{TI} - E_u^{TI}, \tag{2}$$

where the operator $-$ denotes element-wise subtraction. Ideally, if the utterance u comes from the target speaker spk, small values, which are close to a vector of zeros, are expected to be observed in both E_{diff}^{TD} and E_{diff}^{TI}. This is because E_{spk}, a representative of spk's multiple enrolled utterances, is expected to be close in the latent space to the embedding representation of the speaker's test utterance u. If u comes from a non-target speaker, on the other hand, values very different from zeros are expected in both E_{diff}^{TD} and E_{diff}^{TI}. We posit that both E_{diff}^{TD} and E_{diff}^{TI} serve as useful discriminants. Moreover, we expect that E_{diff}^{TD} and E_{diff}^{TI} remain discriminative as users' profiles and test utterances evolve over time. This is because relative small values, close to zeros, remain expected in E_{diff}^{TD} and E_{diff}^{TI} if test utterances come from target speakers.

As detailed in Table 1, E_{spk} and E_u are not always available. In such cases, we cannot calculate the differential embedding directly. To address this issue, when either one or both of the two embedding inputs (i.e., E_{spk} and E_u) from one system are not available, both E_{spk} and E_u from that system will be set to zero vectors. E_{diff} will also be a zero vector as the result of element-wise subtraction. At the same time, we will infer the differential embedding from the other subsystem, that is available. Formally,

$$E_{infer}^{TD} = \begin{cases} 0's, \text{if both } E_{spk}^{TD} \text{ and } E_u^{TD} \text{ are available,} \\ f(E_{diff}^{TI} \cdot W_{TD2TI} + b_{TD2TI}), \text{Otherwise.} \end{cases} \tag{3}$$

Similarly,

$$E_{infer}^{TI} = \begin{cases} 0's, \text{if both } E_{spk}^{TI} \text{ and } E_u^{TI} \text{ are available,} \\ f(E_{diff}^{TD} \cdot W_{TI2TD} + b_{TI2TD}), \text{Otherwise.} \end{cases} \tag{4}$$

where W_{TD2TI}, W_{TI2TD} and b_{TD2TI}, b_{TI2TD} are learnable weight matrices and bias vectors, respectively. $f(\cdot)$ is the activation function. Exponential linear units [15] (ELUs) are chosen as the activation function for their expected benefits in convergence and accuracy. The inferred differential embedding serves as a substitute that we only rely on when the direct differential embedding is not available. Therefore,

$$\hat{E}_{diff}^{TD} = E_{diff}^{TD} + E_{infer}^{TD}. \tag{5}$$

Similarly,

$$\hat{E}_{diff}^{TI} = E_{diff}^{TI} + E_{infer}^{TI}. \tag{6}$$

We then combine the differential information from TD and TI systems by concatenating $\hat{E}_{diff}^{TD}$ and $\hat{E}_{diff}^{TI}$.

$$\hat{E}_{diff} = \hat{E}_{diff}^{TD} \| \hat{E}_{diff}^{TI}. \tag{7}$$

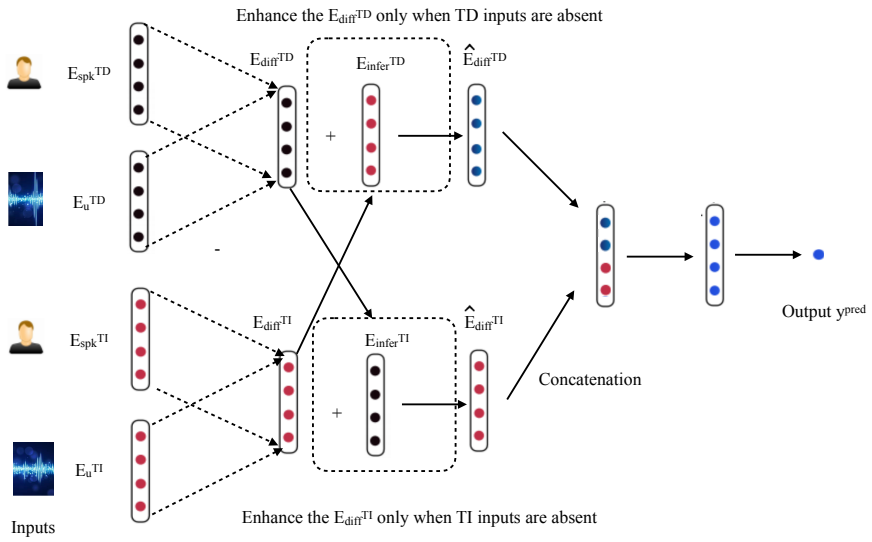

Fig. 1: FOE*net* architecture. Two differential embeddings E_{diff}^{TD} and E_{diff}^{TI} are first calculated based on speaker profile and test utterance embeddings, for TD and TI systems separately. In addition, two backup differential embeddings E_{infer}^{TD} and E_{infer}^{TI} are inferred from TI and TD systems, respectively. The backup differential embeddings are used only when E_{diff}^{TD} or E_{diff}^{TI}, respectively, are not available. $\hat{E}_{diff}^{TD}$ and $\hat{E}_{diff}^{TI}$ combines the information from direct differential embedding and inferred differential embedding. Finally $\hat{E}_{diff}^{TD}$ and $\hat{E}_{diff}^{TI}$ are concatenated and fed to a connected layer.

We further feed $\hat{E}_{diff}$ to a fully connected layer to yield the prediction score y^{pred}. Formally,

$$y^{pred} = f(BN(\hat{E}_{diff} \cdot W_{pred} + b_{pred})), \tag{8}$$

where W_{pred} and b_{pred} are learnable parameters. $BN(\cdot)$ conducts the batch normalizations, and $f(\cdot)$ is the activation function (the sigmoid function in our implementation).

We use cross-entropy as the loss for an instance i:

$$l_i(\Theta) = -[y_i \log(y_i^{\text{pred}}) + (1 - y_i) \log(1 - y_i^{\text{pred}})], \tag{9}$$

where Θ denotes the union of all model parameters.

Given a mini-batch of k training instances, the loss over all k training instances is given by

$$L(\Theta) = \frac{1}{k} \sum_{i=1}^{k} l_i(\Theta) + \alpha L_2(\Theta), \tag{10}$$

where $L_2(\Theta)$ calculates the L_2 norm of Θ as the regularizer and α is a hyperparameter, which balances the trade-off between the classification loss and regularizer. To update the model parameters Θ, we apply Adam [16].

3. EXPERIMENTS

We follow previous work [17] to extract acoustic features from the raw audio. Forty-dimensional Mel-spectrograms are extracted from waveforms after energy-based voice activity detection.

3.1. Training and evaluation data

The training and evaluation are conducted on anonymized speech utterances collected from Alexa devices. The training data is composed of both positive and negative instances. To construct positive instances, we use an existing speaker recognition model and randomly select around 1M utterance/speaker pairs, as identified by high prediction scores in single-speaker registered households. In this way, it is less likely that non-target speakers are included in the positive training instances. The same number of negative instances are then constructed by pairing a random guest speaker and a test utterance from the same gender but from different households. The gender is determined by a pre-trained utterance-level gender classifier. This selection strategy is based on the intuition that confusable (same-gender) pairs are more informative for training. After constructing the dataset, 85% of instances are utilized as parameter training data and the remaining 15% serve as validation data, i.e., for model selection.

The evaluation dataset is constructed by first randomly sampling anonymized utterances. Then each sampled utterance and the enrollment data of speakers associated with the same device are sent to multiple annotators to collect the ground-truth labels independently. To reduce annotation errors, we select utterances which have consistent annotation labels to form the final evaluation dataset. For comparisons over time, we construct two such evaluation datasets based on data from two consecutive weeks.

Table 2: FRR relative improvements in Week 1

Scenarios	Methods	Targeted FAR			
		0.8%	2.0%	5.0%	12.5%
TD&TI present	FOE*net* vs GE2E	21.0%	22.5%	22.8%	22.5%
	FOE*net* vs AF	10.3%	10.3%	14.4%	14.1%
TD absent	FOE*net* vs GE2E	14.8%	19.4%	30.5%	47.7%
	FOE*net* vs AF	17.2%	20.3%	31.7%	49.3%
TI absent	FOE*net* vs GE2E	-	-	-	-
	FOE*net* vs AF	35.3%	40.7%	48.9%	50.1%

Table 3: FRR relative improvements in Week 2

Scenarios	Methods	Targeted FAR			
		0.8%	2.0%	5.0%	12.5%
TD&TI present	FOE*net* vs GE2E	20.5%	21.7%	25.0%	22.5%
	FOE*net* vs AF	8.3%	12.1%	13.1%	14.6%
TD absent	FOE*net* vs GE2E	14.3%	20.6%	30.7%	45.0%
	FOE*net* vs AF	15.2%	20.6%	30.7%	46.0%
TI absent	FOE*net* vs GE2E	-	-	-	-
	FOE*net* vs AF	34.9%	40.7%	49.9%	53.3%

3.2. Baselines

To evaluate the performance of FOE*net*, the following two methods are adopted as baselines:

Average Fusion (AF) takes the prediction scores from TD and TI models as inpus and outputs the average of these two scores as the fused prediction score. When TD or TI score is absent, we construct a piecewise linear score mapping based on pairs of corresponding FAR numbers between AF and the TD/TI system.

Single-system GE2E [18] utilizes an LSTM to construct utterance embeddings and optimizes the speaker recognition system by maximizing the similarity among utterances coming from the same speaker. It utilizes the entire utterance to make predictions and serves as a TI model. No TD system is used.

In this work, the TD model is trained based on the Siamese network [19] and the TI model is trained based on GE2E. We highlight that the proposed embedding fusion framework can be applied to any embedding-based TD and TI systems.

3.3. Recognition Performance

In this section, we evaluate the performance of FOEnet against different baseline methods on the two one-week online evaluation datasets. We calculate false accept rate (FAR) and false reject rate (FRR) on the evaluation sets.

$$\text{FAR} = \frac{\#\text{ of false accepted imposter trials}}{\#\text{ of imposter trials}}. \quad (11)$$

$$\text{FRR} = \frac{\#\text{ of false rejected target trials}}{\#\text{ of target trials}}. \quad (12)$$

As different downstream applications have different requirements for expected FARs, we threshold scores at four different values, and compare corresponding FRR results across methods.

Tables 2 and 3 compare the FRRs for different methods on the two evaluation datasets. Precisely, the relative FRR reduction between FOEnet and the other two baselines are calculated. We highlight three main observations. First, we find that both FOEnet and AF outperform GE2E, showing the advantage of utilizing both TD and TI models in combination. Leveraging both TD and TI systems yields more accurate speaker identity predictions than just depending on a single TI model.

Second, FOEnet consistently outperforms the single system (GE2E) and the fusion-by-average (AF) baselines. For example, when we target 0.8%, 2.0%, 5.0%, and 12.5% FARs, compared with GE2E/AF, FOEnet reduces FRR by 21.0%/10.3%, 22.5%/10.3%, 22.8%/14.4%, and 22.5%/14.1%, respectively, when both TD and TI inputs are available in Week 1. This observation also applies to the evaluation dataset from Week 2. We conclude that fusion at the embedding level, rather than at the score level, significantly improves the recognition performance.

Third, for almost all FAR levels, the advantage of FOEnet over GE2E and AF increases when there are incomplete inputs (note that GE2E can only be applied when TI input is present). Specifically, when TD signals are not available, FOEnet outperforms GE2E and AF by 47.7% and 49.3% respectively at a target FAR of 12.5% (in Week 1). This shows that our approach of inferring missing differential embeddings prior to system fusion is effective.

We also confirmed the intuition that generating negative samples based on same-gender pairings improves model performance. There is a relative 6% FRR reduction when gender information is used for ruling out easy, different-gender samples. Due to the space limit, detailed numbers regarding FOEnet vs FOEnet w/o same-gender pairings are not presented.

4. CONCLUSIONS

We presented FOEnet, a fusion architecture that combines predictions from both TD and TI speaker recognition systems at the embedding-level, by combining differential embeddings from both subsystems. Special care is taken to allow the system to work when either TD or TI input is not available, as it may happen due to the operational characteristics of voice assistants (such as missing wakewords). This is accomplished by estimating the missing differential embeddings from the embeddings that are available. FOEnet does not require fine tuning of the previously trained TI and TD models; both models can be directly used as is. Experiments on Alexa voice assistant traffic data demonstrate that FOEnet is substantially more effective than either a single TI GE2E model or a score-averaging fusion method.

5. REFERENCES

[1] Hatice Gunes and Massimo Piccardi, "Affect recognition from face and body: early fusion vs. late fusion," in *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, Waikoloa, Hawaii, USA, October 10-12, 2005*. 2005, pp. 3437–3443, IEEE.

[2] Cees Snoek, Marcel Worring, and Arnold W. M. Smeulders, "Early versus late fusion in semantic video analysis," in *Proceedings of the 13th ACM International Conference on Multimedia, Singapore, November 6-11, 2005*, HongJiang Zhang, Tat-Seng Chua, Ralf Steinmetz, Mohan S. Kankanhalli, and Lynn Wilcox, Eds. 2005, pp. 399–402, ACM.

[3] Najim Dehak, Patrick Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-End Factor Analysis for Speaker Verification," *IEEE Trans. Audio, Speech & Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[4] Chao Li, Xiaokong Ma, Bing Jiang, Xiangang Li, Xuewei Zhang, Xiao Liu, Ying Cao, Ajay Kannan, and Zhenyao Zhu, "Deep Speaker: an End-to-End Neural Speaker Embedding System," *CoRR*, vol. abs/1705.02304, 2017.

[5] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "VoxCeleb2: Deep Speaker Recognition," *CoRR*, vol. abs/1806.05622, 2018.

[6] Weidi Xie, Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "Utterance-level Aggregation for Speaker Recognition in the Wild," in *Proceedings of ICASSP, Brighton, United Kingdom, May 12-17*, 2019.

[7] Mahdi Hajibabaei and Dengxin Dai, "Unified Hypersphere Embedding for Speaker Recognition," *CoRR*, vol. abs/1807.08312, 2018.

[8] Feng Wang, Weiyang Liu, Hanjun Dai, Haijun Liu, and Jian Cheng, "Additive Margin Softmax for Face Verification," in *Proceedings of ICLR, Vancouver, BC, Canada, April 30 - May 3*, 2018.

[9] Andy T. Liu, Shu-Wen Yang, Po-Han Chi, Po-chun Hsu, and Hung-yi Lee, "Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders," in *Processing of ICASSP 2020, Barcelona, Spain, May 4-8, 2020*. IEEE.

[10] Andy T. Liu, Shang-wen Li, and Hung-yi Lee, "TERA: self-supervised learning of transformer encoder representation for speech," *CoRR*, vol. abs/2007.06028, 2020.

[11] Sachin S. Kajarekar, Nicolas Scheffer, Martin Graciarena, Elizabeth Shriberg, Andreas Stolcke, Luciana Ferrer, and Tobias Bocklet, "The SRI NIST 2008 speaker recognition evaluation system," in *Proc. IEEE ICASSP*, 2009, pp. 4205–4208.

[12] Lukáš Burget, Michal Fapšo, Valiantsina Hubeika, Ondřej Glembek, Martin Karafiát, Marcel Kockmann, Pavel Matějka, Petr Schwarz, and Jan Černockỳ, "BUT system for NIST 2008 speaker recognition evaluation," in *Proc. Interspeech*, 2009, pp. 2335–2338.

[13] Suwon Shon, Tae-Hyun Oh, and James R. Glass, "Noise-tolerant audio-visual online person verification using an attention-based neural network fusion," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*. 2019, IEEE.

[14] Iosif Mporas, Saeid Safavi, and Reza Sotudeh, "Improving robustness of speaker verification by fusion of prompted text-dependent and text-independent operation modalities," in *Speech and Computer - 18th International Conference, SPECOM 2016, Budapest, Hungary, August 23-27, 2016, Proceedings*, Andrey Ronzhin, Rodmonga Potapova, and Géza Németh, Eds. 2016, vol. 9811 of *Lecture Notes in Computer Science*, pp. 378–385, Springer.

[15] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," in *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

[16] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun, Eds., 2015.

[17] Prashant Anand, Ajeet Kumar Singh, Siddharth Srivastava, and Brejesh Lall, "Few Shot Speaker Recognition using Deep Neural Networks," *CoRR*, vol. abs/1904.08775, 2019.

[18] L. Wan, Q. Wang, A. Papir, and I. Moreno, "Generalized End-to-end Loss For Speaker Verification," in *Proceedings of ICASSP, Calgary, AB, Canada, April 15-20*, 2018, pp. 4879–4883.

[19] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah, "Signature verification using a siamese time delay neural network," in *Advances in Neural Information Processing Systems, [7th NIPS Conference, Denver, Colorado, USA, 1993]*, 1993, pp. 737–744.