BRIDGING TEXT DEPENDENT WITH TEXT INDEPENDENT SPEAKER IDENTIFICATION BY EMBEDDING FUSION

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

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

ABSTRACT

Speaker identification answers the fundamental question "Who is speaking?". The identification technology enables various downstream applications such as providing a personalized experience and proffering expeditious shopping checkouts. Based on whether the speech content is constrained or not, there are text-dependent and text-independent speaker identification models. To yield more accurate speaker identification and better serve customers, text-dependent and text-independent models are expected to be ensembled to make predictions, which are more reliable.

We highlight that a good speaker identification system is expected to not only yield accurate predictions with complete inputs from dependencies, but also consistently perform well even with incomplete inputs for robustness. To leverage predictions from multiple systems and alleviate the potential problems caused by input data incompleteness, we propose an Embedding Fusion Network method (*EFN*) for speaker identification. *EFN* emphasizes collaborative learning to recognize speakers with attention to fuse predictions from multiple systems and handling input incompleteness. We conduct experiments on an Alexa dataset and compare *EFN* with two state-of-the-art baseline methods. The results conclude that *EFN* achieves higher accuracy than all baseline methods, especially when there are incomplete inputs.

1. INTRODUCTION

Speaker identification answers the fundamental question "who is speakeing" for a speech utterance. Based on whether the utterance content is constrained or not, there are two types of speaker identification models, i.e., text-dependent (TD) models and text-independent (TI) models. For the first one, the utterance content is fixed, usually based on the predefined wakeword, such as "Hey Google","Echo", or "Alexa". For the latter one, there are no content requirements of an utterance. TD speaker identification aims to achieve expeditious speaker recognition while TI speaker identification strives to utilize the entire utterance to provide accurate recognition.

Given an utterance, for example, "Alexa, what is the time now?" the TD and TI speaker identification models can make speaker identity predictions independently. On the one hand, the TD model compares the acoustic signal between "Alexa" and a user's corresponding enrolled wakeword segments to calculate a matching score. On the other hand, the TI model can take either the entire utterance or the non-wakeword segment and compare it with a user's enrolled utterances to yield a matching score. The two models are generally trained with independent machine learning models and on different datasets to achieve the best identification performances, respectively. With both TD and TI models, how to effectively leverage predictions from each individual model to predict the speaker identity of a test utterance remains a daunting task due to the following two concerns. First, it is not clear how to leverage TD and TI speaker identification models in a collaborative manner to serve customers. Second, the inputs from the two systems, especially the TD system, may not be always available as it is not necessary for an utterance to contain a wake word. Table 1 shows the scenarios when there will be missing utterance or speaker profile embeddings. Therefore, it is challenging to proffer comparable speaker identification predictions with multiple models when facing incomplete model inputs.

Speaker identification has been studied as a supervised classification problem based on the characteristics of voices. Traditionally, the prevalent solution is based on i-vector representation of speech segments [1]. Recently, deep learning-based methods have been gaining attention as they outperformed the prevalent i-vector solutions. Deep speaker [2] utilizes residual networks to extract audio embeddings and optimizes triplet loss to perform speaker identification. VGGVox [3] adopts a CNN-based residual network and optimizes contrastive loss to train the model. Two other Resnet-based methods utilize additive margin softmax classification loss [4] to improve the recognition accuracy in [5, 6]. More recently, Mockingjay [7] and TERA [8] learn speaker representations through spectrogram reconstruction based on self-supervised training. However, very few work investigate leveraging multiple speaker identification systems to improve predication accuracy and robustness. [9] explores an online audio-visual fusion system for person verification using face and voice. [10] investigates GMM-UBM based TD and TI speaker identification models. An extra classifier is introduced by feeding the concatenation of TD and TI scores as the inputs. However, it cannot handle the issue of incomplete inputs.

Our work presumes users have a registered profile which consists of four to ten utterances and used to create an embedding representing the enrolled user. There are pretrained TD and TI models, which are capable of generating corresponding fix-length embeddings given an utterance. Unlike

Table 1: Scenarios when speaker profile or utterance embeddings are unavailable during serving

Scenarios	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	✓	✓	✓
A new wakeword is just released to customers	x	x	✓	✓
TD model is working improperly during run time	✓	x	✓	✓
TI model is working improperly during run time	✓	✓	✓	x

The detailed descriptions regarding E_{spk}^{TD} and E_{u}^{TD},E_{spk}^{TI} and E_{u}^{TI} are available in Section 2.1.

the aforementioned fusion research which presumes that complete inputs from different models are always available, we develop a fusion approach targeting to leverage embedding-level predictions from TD and TI models even with incomplete inputs. We foresee the need of fusing various models with incomplete inputs as (1) an utterance may not always have a wakeword. (2) the wakeword detector may fail to identify the wakeword in an utterance. (3) a user haven't get enrolled for a wakeword but the testing utterances involves the wakeword. and (4) run-time identification engines are working improperly temporally and no utterance embedding are generated during serving.

Our proposed solution fuses predictions from a TD model and a TI model on the embedding level with the focus on offering an accurate and robust speaker identification fusion model. We developed an embedding fusion network to fuse the predictions from TD and TI systems (Section 2.1). Our experiments demonstrate that our proposed solution, *EFN*, significantly outperforms the all baseline methods, especially when there are incomplete inputs (Section 3.3), which improves the system robustness.

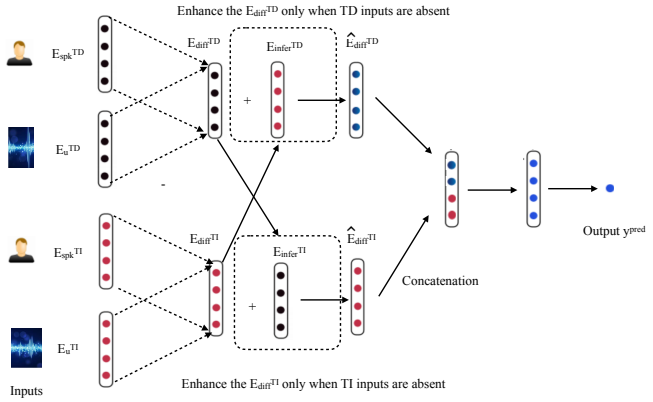

Fig. 1: The framework of *EFN*. The embedding difference between speaker profile embedding and testing utterance embedding is first calculated for both TD and TI systems. At the same time, a backup embedding difference is inferred from the other system and the backup embedding difference is only used when E_{diff} is not available. $\hat{E}_{diff}^{TD}$ and $\hat{E}_{diff}^{TI}$ combines the information from embedding difference and inferred embedding difference. $\hat{E}_{diff}^{TD}$ and $\hat{E}_{diff}^{TI}$ are further concatenated as a guidance to train the remaining fully connected layers.

2. PROBLEM STATEMENT & METHODOLOGY

We formulate the objective of our work as the following. Suppose the system has a set of existing users with registered speech utterances as enrollment data. Given another short testing speech utterance, the goal of this study is to recognize the speaker identity behind the testing speech utterance. Moreover, we focus on how to effectively leverage predictions from both TD and TI models, especially when there are incomplete inputs from either TD or TI models.

To better explain how to fuse predictions from TD and TI models and how to handle incomplete inputs, we illustrate the framework of *EFN* in Fig. 1.

2.1. Embedding Fusion Networks

Given a testing utterance u, the goal of speaker identification model is to learn a scoring function such that the score $s(spk, u)$ from the ground-truth speaker is as high as possible while the scores from other speakers are as low as possible.

We use E_{u}^{TD} and E_{u}^{TI} to denote the embeddings of a testing utterance u generated by a TD and TI model, 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 $E_{spk} = \frac{1}{M}\sum_{i=1}^{M} E_{u_i}$. Please note that the utterance and speaker profile embeddings may not be always available in practice.

To effectively fuse the predictions from both TD and TI models, as well as handling the issue of input incompleteness, we develop a fusion model which leverages predictions on embedding level. When the input from one system is absent, the model will automatically infer the missing signal from the other system, which improves the robustness of the entire speaker identification system.

For both TD and TI models, we first calculate the embedding difference 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

$$E_{diff}^{TI} = E_{spk}^{TI} - E_{u}^{TI}, \quad (2)$$

where the operator $-$ refers to element-wise subtraction. Ideally, if the utterance u comes from 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 that E_{spk}, which is a representative of spk's multiple enrolled utterances, is expected to be close to the embedding representation of his/her another utterance u in the latent space. On the other hand, if u comes from a different speaker other than the ground-truth speaker, relatively large values off a vector of zeros are expected to be observed in both E_{diff}^{TD} and E_{diff}^{TI}. We posit that both E_{diff}^{TD} and E_{diff}^{TI} will serve as an discriminative indicator, which effectively expedites training and distinguish genuine speakers from others.

To handle the missing input issues as shown in Table 1, 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 will be set to zero vectors. As a result, E_{diff} will be a zero vector. In such cases, we will infer the difference vector from the other system. Formally,

$$E_{infer}^{TD} = \begin{cases} 0's, \text{if either } E_{spk}^{TD} \text{ or } E_{u}^{TD} \text{ is available,} \\ f(E_{diff}^{TI} \cdot W_{TD2TI} + B_{TD2TI}), \text{Otherwise.} \end{cases} \quad (3)$$

Similarly,

$$E_{infer}^{TI} = \begin{cases} 0's, \text{if either } E_{spk}^{TI} \text{ or } E_{u}^{TI} \text{ is available,} \\ f(E_{diff}^{TD} \cdot W_{TI2TD} + B_{TI2TD}), \text{Otherwise.} \end{cases} \quad (4)$$

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 Unit [11] (ELU) is chosen here because of the expected benefits in convergence and accuracy. The inferred embedding difference serves as a backup and we only rely on the inferred embedding difference when its own inputs are not available. Therefore,

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

Similarly,

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

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

$$\hat{E}_{diff} = \hat{E}_{diff}^{TD} \| \hat{E}_{diff}^{TI}. \quad (7)$$

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

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

where W_{pred} and B_{pred} are learnable parameters. $BN(\cdot)$ conducts the batch normalizations. and $f(\cdot)$ is the activation function, where the $Sigmoid$ function is utilized here.

We apply the cross-entropy to calculate the loss for an instance i:

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

where Θ denotes the model parameters.

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

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

where $L_2(\Theta)$ calculates the $L2$ norm of Θ as the regularizer and α balances the trade-off between the classification loss and the regularizer.

To update the model parameters, we apply Adam [12] to optimize Θ. Algorithm 1 summarizes the detailed training processes of *EFN*.

Algorithm 1: Embedding fusion model training

Input: learning rate η, maximal number of epochs itr_{max}, TD speaker embedding E_{spk}^{TD}, TD utterance embedding E_{u}^{TD}, TI speaker embedding E_{spk}^{TI}, TI speaker embedding E_{spk}^{TI}

Initialize: Model parameters Θ, EER* = $+\infty$

/* Training on the training dataset */
for $itr \leq itr_{max}$ **do**
  Randomly shuffle all training instances.
  Sequentially generate a mini-batch of k training instances.
  Calculate loss based on Equation 10.
  Update model parameters by applying Adam.

/* Calculate the EER on the validation dataset */
if EER $\leq$ EER* **then**
  Θ^* = Θ
  EER* = EER

Output: Θ^*

3. EXPERIMENTS

In this work, We follow the previous work [13] to extract acoustic features from the raw audios. 40 dimensional LFBE features are extracted from speech after an energy-based voice activity detection.

3.1. Training and evaluation data

The training and evaluation are conducted on real speech utterances collected from Alexa devices. We first discuss the construction of the training data and subsequently explain how to form the evaluation data. To construct positive instances in training data, we use an existing speaker identification model and randomly select around 900,000 utterance speaker pairs, which has high predication scores in single registered speaker households. In this way, it is less likely that non-ground-truth speakers are involved in the positive instances, which guarantees the quality of the positive training data. To construct informative negative training instances, we further leverage the intuition that having hard negative examples results in better accuracy. This is done through a

pre-trained utterance-based gender classifier. The same number of negative instances are then constructed by pairing a candidate speaker and a testing utterance from the same gender. After constructing the training dataset, 85% of instances are utilized as the training data and the remaining 15% of instances serve as the validation data, which helps select the best training model.

To construct the evaluation dataset, we randomly sample around 10,000 utterances from all the traffic in a week. Then each sampled utterance and the enrolled utterances of the involved speakers under the same device are sent to three annotators for collecting the ground-truth labels. To reduce the annotation errors, we select those utterances which have at least two consistent annotation labels as the evaluation data. To make persistent comparisons, we construct two such evaluation datasets based on two consecutive weeks.

3.2. Baselines

To evaluate the performance of *EFN*, the following two methods are adopted as baselines.

- **Average Fusion (AF)** takes the predication scores from TD and TI models as the input and output the average of these two scores as the fused prediction score.
- **GE2E** [14] uses an LSTM-RNN to construct utterance embeddings and optimizes the speaker identification 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 systems are involved.

In this work, the TD model is trained based on the Siamese network [15] and the TI model is trained based on GE2E. But the proposed embedding fusion framework can be generalized and applied to other TD and TI systems.

3.3. Recognition Performance

In this section, we evaluate the performances of *EFN* against different baseline methods on two 1-week online evaluation datasets. We calculate false accept rate (FAR) and false reject rate (FRR) on the evaluation data.

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

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

As different downstream applications have different requirements on expected FAR or FRR. In this work, we threshold at different identification score to accept/reject candidate speakers. More precisely, we compare the FRR performances among different methods at 4 different targeted FARs.

Tables 2 and 3 show the FRR performance of different methods on the two evaluation datasets. We have three observations from the results. First, EFN consistently outperforms GE2E and AF. For example, when we target 0.8%, 2.0%, 5.0%, and 12.5% FARs, compared with GE2E/AF, EFN 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 one. This observation also applies the evaluation dataset in week two. The performance gain stems from the effective collaborative learning from TD and TI models on the embedding level. The predictions from one system can complement the predictions from the other and improve the overall identification performance.

Table 2: FRR relative improvements in week one.

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

Table 3: FRR relative improvements in week two.

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

Second, EFN is more effective when there are incomplete inputs. For example, when TD signals are not available, EFN outperforms GE2E and AF by 47.7% and 49.3% respectively at targeted FAR 12.5% in week one. The rationale behind it is when there are missing inputs from one system, it is more reliable to infer the missing information from the other system. These inferred signals yield more accurate speaker identifications, which enhances the robustness of the system.

Third, the comparison also indicates that EFN and AF outperform GE2E, which shows the advantages of utilizing both TD and TI models to serve customers, which yields more accurate identity predictions than a pure TI model does.

We also discover that negative sampling based on the same gender improves the model performance. There is relative 6% FRR improvement across different targeted FARs when gender information is used for choosing harder negative samples. Due to the space limit, detailed results are not presented in the table.

4. CONCLUSION

We presented *EFN*, a framework that combines predictions from both TD and TI speaker identification systems at the output embedding-level with the focus on providing a robust speakerID system even when facing incomplete model inputs under different scenarios. Our method does not require fine tuning of the already optimized TI and TD models and both models can be used out of the box. Our experiments on real customer datasets show that the proposed solution outperforms the two baseline methods: average fusion and GE2E.

5. REFERENCES

[1] 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.

[2] 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.

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

[4] 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.

[5] 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, pp. 5791–5795.

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

[7] 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 *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*. 2020, pp. 6419–6423, IEEE.

[8] 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.

[9] 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, pp. 3995–3999, IEEE.

[10] 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.

[11] 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*, Yoshua Bengio and Yann LeCun, Eds., 2016.

[12] 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.

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

[14] 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.

[15] 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]*, Jack D. Cowan, Gerald Tesauro, and Joshua Alspector, Eds., 1993, pp. 737–744.