Multimodal Association for Speaker Verification

Suwon Shon[1], *James Glass*[2]

[1]ASAPP Inc., New York, NY, USA
[2]MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, USA
{swshon,glass}@csail.mit.edu

Abstract

In this paper, we propose a multimodal association on a speaker verification system for fine-tuning using both voice and face. Inspired by neuroscientific findings, the proposed approach is to mimic the unimodal perception system benefits from the multisensory association of stimulus pairs. To verify this, we use the SRE18 evaluation protocol for experiments and use out-of-domain data, Voxceleb, for the proposed multimodal fine-tuning. Although the proposed approach relies on voice-face paired multimodal data during the training phase, the face is no more needed after training is done and only speech audio is used for the speaker verification system. In the experiments, we observed that the unimodal model, i.e. speaker verification model, benefits from the multimodal association of voice and face and generalized better than before by learning channel invariant speaker representation.

Index Terms: speaker verification, fine-tuning, multimodal, SRE18

1. Introduction

In recent years, deep neural network (DNN) based speaker embeddings have become the state-of-the-art approach for the task of speaker verification. Various DNN architectures and loss functions have dramatically boosted system performance. Traditionally, the series of Speaker Recognition Evaluations (SRE) hosted by the National Institute of Standards and Technology (NIST) enable researchers to benchmark their system easily by providing appropriate dataset and metrics. While NIST consistently provided an 8kHz telephony speech as the basic audio format, the channel domain mismatch problem was the main issue for recent evaluations and many participants tried to solve the problem by proposing an domain adaptation or compensation algorithm. Unfortunately, the SRE dataset was only available to researchers who fully participated in the event and it was difficult to access or be investigated by new researchers.

Independently, the Voxceleb [1, 2] dataset was recently released based on the excellent performance of face recognition technology. This publicly available dataset accelerated investigation on speaker verification by allowing anyone who wanted to study speaker verification technology. Unfortunately the dataset contained artifacts due to the nature of the recordings. Voxceleb was collected only from celebrities to avoid violating individual privacy rights. However, celebrities could be categorized by their specialty, which tended to have a strong correspondence with the visual and auditory background. For example, sports athletes appeared in crowded and noisy environments while newscasters were generally found in a relatively calm studio with high-quality audio. This recording domain artifact enabled DNN-based speaker verification embeddings to easily overfit on the training data, since they would learn the background recording information to verify someone's voice.

Both channel domain mismatch and recording domain match can be regarded as overfitting issues for DNN-based speaker verification. To address the problem, many studies report effective algorithms that actively use the channel domain label by categorizing the channel or by adversarially training the DNN model to prevent the utilization of information from the channel.

In this paper we propose a new approach to prevent overfitting and to generalize the DNN model to make the speaker verification system operate well on the target domain. The main concept of the proposed approach is the fine-tuning of the DNN model. Fine-tuning generally uses the data of the target domain to be applied, and it is very easy to fall into the overfitting problem. However, we propose a method for fine-tuning by using the data of completely different domains, even from different modalities, without using the data of the target domain. This approach was motivated by the observation of a neuroscience study that unimodal perception benefits from the multisensory association of ecologically valid and sensory redundant stimulus pairs [3]. Note that this approach is investigated for speech-based speaker verification combined with another modality, such as visual facial information.

2. Related work

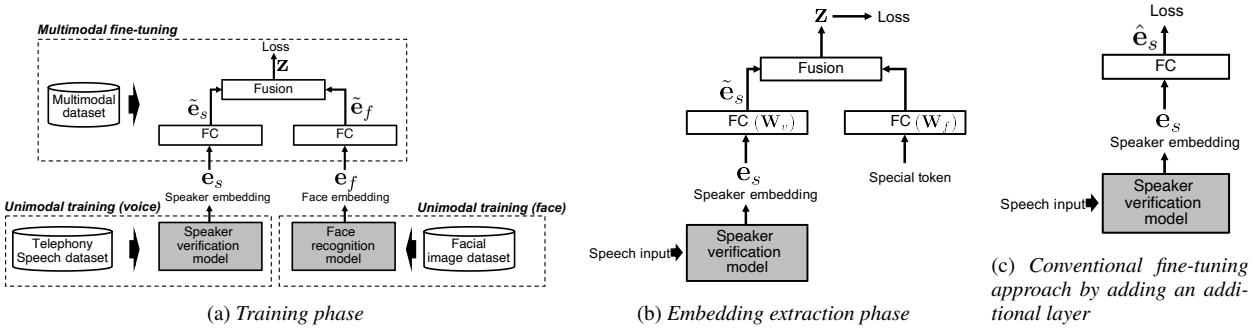
In this section, we describe the relevant studies related to the problem we will address and also describe the fine-tuning techniques for the speaker verification system that we will use.

2.1. Channel domain mismatched condition

A channel domain mismatched condition occurs when the channel domain of the training data is not matched with the target (i.e. test) domain. For example, if the system was trained using only speech data collected from YouTube, it does not guarantee that the system will operate the same when the speech came from different sources like telephone or interview recordings, even if they have the same audio file specification format. It can be also said a system is overfit on YouTube. This situation tends to arise due to a lack of various data from different channels. For an i-vector-based speaker verification system, a solution can be achieved if the source domain data has a channel label [4, 5, 6] or a small set of target domain data is available [7, 8, 9, 10, 11, 12, 13]. For DNN-based speaker embedding speaker verification, combining multiple losses for training with an additional label of noise or condition is also popular [14, 15, 16, 17, 18, 19, 20].

2.2. Recording domain matched condition

The recording domain matched condition is a relatively new issue to be investigated because celebrity voices were not generally used to train a speaker verification system. As mentioned



(a) *Training phase* (b) *Embedding extraction phase* (c) *Conventional fine-tuning approach by adding an additional layer*

Figure 1: *High-level framework for multimodal fine-tuning approach.* $\mathbf{e}_s$*: speaker embedding,* $\mathbf{e}_f$*: face embedding. Note that both* $\tilde{\mathbf{e}}_s$ *and* $\mathbf{z}$ *can be generated only using speech on embedding extraction phase.*

in the previous section, celebrities have a high correlation with the recording environment of their respective discipline. Recent studies show that if we utilize self-supervision of the dataset to provide information to the network whether two speech segments came from the same recording or a different recording, the network can generate channel invariant speaker embeddings by removing recording or background environment information that has a high correlation with individual celebrities [21, 22]. While these studies succeeded to encourage networks not to incorporate recording domain information by combining adversarial loss with existing cross-entropy loss, they failed to show it also operates well on the channel domain mismatches condition such as telephony channel.

2.3. Fine-tuning for speaker verification

Fine-tuning of the DNN model can be done to a pre-trained model that is trained for the same or different tasks. If the dataset for fine-tuning has no variability on the channel domain and only contains target domain data, then the system will end up with an over-fit model. At the same time, this over-fit model would also guarantee the reasonable performance on the target domain [23]. For fine-tuning speaker verification systems, many different approaches have been presented such as using different losses compared to the pre-trained model, adding an additional layer, and updating part or all of the layers [1, 2, 24, 25, 26] Not surprisingly, all these methods are using the matched target domain data for fine tuning.

3. Multimodal fine-tuning on speaker verification system

The main idea of the proposed approach is fine-tuning of the pre-trained model to improve the performance for the task of speaker verification. However, if we use the target domain data for fine-tuning, the model cannot be free from over-fitting. Thus, we use a multimodal dataset collected from a completely different domain for fine-tuning. Here, we assume that the fine-tuned DNN model generalizes well if the model performs better on the target domain using the fine-tuning data collected from another domain rather than the target domain.

3.1. Motivation

Our approach is motivated by a neuroscience study about implicit multisensory associations of the human perception system [3]. In the study, it was observed that unimodal perception benefits from the multisensory association of ecologically valid and sensory redundant stimulus pairs. In particular, the study showed that human participants became better on a speaker verification task after exposure to a paired face which, the authors claim, induces a multisensory association. To mimic this effect, we will use a YouTube video dataset that has a large amount of voice and face pairs for fine-tuning of a telephony channel speaker verification system. Note that the telephony channel domain rises in a significantly different acoustic channel compared to YouTube videos and this channel mismatched condition causes a large performance gap. However, even if the multimodal dataset is out-of-domain to the target, we hypothesized that the association of the face and voice would make a positive impact on the unimodal system, i.e speaker verification model, as was found in the neuroscience study. We observed a similar effect for multimodal person verification when one of the modality is incomplete from a previous study [27].

3.2. Unimodal models

To verify the proposed approach, we need two pre-trained models, a speaker verification model, and a face recognition model. The two models have a similar DNN architecture. Once they have speech audio or a facial image as input, convolutional neural network (CNN) layers produce a 3 dimensional output representation. This representation can be concatenated or averaged to feed into a fully connected layer. The loss function is applied after the fully connected layer. The loss can be cross-entropy loss with softmax, triplet loss, or contrastive loss using speaker or face identity labels. A noticeable difference between the two models is that the audio has variable length while the face image size is fixed. Thus, the concatenation type of pooling cannot be used to aggregate the CNN layer output for the speaker verification model.

3.3. Multimodal fine-tuning

Based on the unimodal models, we added a few additional layers on top of the models for fusion. This allows the new additional layer to learn identity information for each modality by associating the multimodal inputs. The high-level framework is shown in Figure 1. We didn't consider updating pre-trained unimodal models to prevent catastrophic forgetting about telephony speech. For fusion, we used the weighted sum of the two embeddings ($\tilde{\mathbf{e}}_s$ and $\tilde{\mathbf{e}}_f$). This is the same approach in our previous study [27]. The weight is calculated by the attention layer. Contrastive loss was used for fine-tuning.

Although we were motivated by the neuroscience study, we still want to understand how the other modality, i.e. face image,

affects the unimodal models. For simplicity, suppose the fusion is score level fusion which is sum of the distance of each embeddings. Then the output of fusion is $D = \|\tilde{\mathbf{e}}_{s,i}, \tilde{\mathbf{e}}'_{s,i}\|_2 + \|\tilde{\mathbf{e}}_{f,i}, \tilde{\mathbf{e}}'_{f,i}\|_2$ where $\|\cdot\|_2$ denotes the l_2 norm of a vector, $(\tilde{\mathbf{e}}_{s,i}, \tilde{\mathbf{e}}'_{s,i})$ and $(\tilde{\mathbf{e}}_{f,i}, \tilde{\mathbf{e}}'_{f,i})$ are the pairs of training speech and face data, respectively. $Y_i \in \{0, 1\}$ and $Y_i = 0$ for the same identity, i.e. same person, and $Y_i = 1$ for the different identity. Let $W_s^{(j,k)}$ is the j-th row and k-th column element of the fully connected layer weight parameter matrix $\mathbf{W}_s \in \mathbb{R}^{h \times d}$ for speaker embedding. By the chain rule, the gradient of the contrastive loss function L is given by

$$\frac{\partial L}{\partial W_s^{(j,k)}} = \frac{\partial L}{\partial D} \frac{\partial D}{\partial W_s^{(j,k)}} \tag{1}$$

where the contrastive loss function $L = (1 - Y_i)\frac{1}{2}(D)^2 + (Y_i)\frac{1}{2}\{max(0, m - D)\}$ and m is the margin. Suppose $Y_i = 0$, then the partial derivative $\frac{\partial L}{\partial D} = D$ and Equation (1) is as,

$$\begin{aligned}
\frac{\partial L}{\partial W_s^{(j,k)}} &= D \frac{\partial D}{\partial W_s^{(j,k)}} \\
&= (\|\tilde{\mathbf{e}}_{s,i}, \tilde{\mathbf{e}}'_{s,i}\|_2 + \|\tilde{\mathbf{e}}_{f,i}, \tilde{\mathbf{e}}'_{f,i}\|_2) \frac{\partial D}{\partial W_s^{(j,k)}} \\
&= \underbrace{\|\tilde{\mathbf{e}}_{s,i}, \tilde{\mathbf{e}}'_{s,i}\|_2 \frac{\partial D}{\partial W_s^{(j,k)}}}_{\text{Gradient by speech}} + \underbrace{\|\tilde{\mathbf{e}}_{f,i}, \tilde{\mathbf{e}}'_{f,i}\|_2 \frac{\partial D}{\partial W_s^{(j,k)}}}_{\text{Gradient by face}}
\end{aligned} \tag{2}$$

As shown in the equation, the partial derivative is divided into two parts with distances from each modality. Consequently, the speech side weight parameter $W_s^{(j,k)}$ is updated by not only audio but also by visual input. In this way, we expect that the network could learn exclusive information from another modality and the model would generalize better by exposing the multimodal association.

A similar study was done on multimodal learning using a deep autoencoder [28]. They proposed a shared representation that can be used for both audio and video construction. However, the study was mainly focused on learning the explicit relationship between lip and speech at the frame level for better video-only feature representation. Thus, it is difficult to verify the benefit of the implicit multisensory association from their study.

4. Experiments

4.1. Evaluation condition

NIST regularly conducts SRE events to evaluate the state-of-the-art of speaker verification techology. Prior to 2018, NIST mainly considered telephone conversational speech for evaluation tasks. In 2018 (SRE18), NIST included a new condition using speech from amateur internet videos to cover speech from various recording channels. In this study, we only evaluate the performance of the speaker verification system on telephone conversational speech, i.e. Call My Network 2 (CMN2). For the channel mismatched condition, we will use speech from the amateur internet videos, i.e. YouTube, as a training dataset for the proposed multimodal fine-tuning approach. The CMN2 dataset consists of development and evaluation sets. We used the CMN2 development set as a validation set for DNN training. The CMN2 evaluation set was not used for any kind of training of parameters or hyper-parameters in the speaker verification system. Performance measurement was done using Equal Error Rate (EER) and minimum detection cost function (C_{min}) that was defined for SRE18.

4.2. Training details

An x-vector system [29] was used for the speaker verification model. A total of 359,463 utterances from 11,900 speakers from SRE 04,05,06,08,10, Mixer-6, voxceleb1 and 2 development sets, and switchboard dataset wre used for training. For the Voxceleb 1 and 2 datasets, utterances from the same audio file were concatenated into a single wav file. Augmentation using the MUSAN [30] noise dataset was also done on the training set and randomly selected 150k utterances. Utterances shorter than 5 seconds after Voice Activity Detection (VAD) were removed. Utterances from speakers that have fewer than 8 utterances were also filtered out from the training. We used 8Khz audio as input, and any audio with a sampling rate higher than 8kHz was downsampled to 8kHz.

We extracted 23 dimensions MFCC from 8kHz audio and a simple energy-based VAD was applied to remove silence. We used two different x-vector architectures from the previous study [31]. The first one is a TDNN model as implemented in the Kaldi [32] egs/SRE16 Recipe. The other one is an E-TDNN architecture which has a slightly wider temporal context of the TDNN and interleaving dense layers in between the convolutional layers. This architecture has been found to greatly outperform the baseline TDNN model on the SRE18 benchmark [31]. For the backend processing after x-vector extraction, we use LDA and PLDA as the general procedure for the x-vector system [29]. We summarized the dataset used for the training of the x-vector system as shown in Table 1.

Table 2 shows the baseline performance of the speaker verification systems on the SRE18 CMN2 condition. Both x-vector systems that were used in this study show a reasonable baseline performance compared to other previous studies. Note that we did not use any score normalization, calibration, or another fine-tuning method that boosts the performance on the x-vector system.

Table 1: *Dataset usage for experimentation.*

	Dataset	Channel domain
x-vector Training (NN)	SRE 04,05,06,08,10	Telephone, microphone
	Voxceleb1-2	YouTube speech
	MUSAN	Various
	Mixer-6	Telephone, microphone
Training(LDA,PLDA)	SRE-Telephone	Telephone
Fine-tuning	SRE-Telephone	Telephone
Multimodal Fine-tuning	Voxceleb 2	YouTube video
Evaluation	SRE18 CMN2 Development and Evaluation sets	Telephone

Table 2: *SRE18 evaluation on CMN2 condition.*

System	EER(%)	
	Development	Evaluation
x-vector (TDNN, $\mathbf{e}_s$)	6.64	7.44
x-vector (E-TDNN, $\mathbf{e}_s$)	5.80	6.57
UTD-CRSS [33] (single best)	7.20	8.63
ViVoLAB [34] (fusion)	-	7.63
DKU-SMIIP [35] (single best)	6.03	6.20
NEC-TT [36] (single best)	-	6.05
JHU-MIT [31] (single best)	4.55	4.95

For face recognition, we used the FaceNet [37] model pre-trained on VGGFace-2 dataset[1]. Since the provided face region annotations in the VoxCeleb datasets are coarse, we re-align and crop faces by the face and landmark detectors in Dlib[2].

For the proposed multimodal fine-tuning, we use the same framework in Figure 1(a). For the fusion, we followed the same network that was proposed in our previous study [27]. Both speaker embedding and face embeddings were extracted in 512 dimensions using the speaker verification and face recognition models. For speaker embeddings, we used the entire utterance as input while the face embedding was extracted from the face in the first frame of each video. Both embeddings were L2-normalized before being fed into the fusion network. We used the Voxceleb 2 development set for training the fusion network. After training was done, the face recognition module was removed and two types of embeddings, $\tilde{\mathbf{e}}_s$ and $\mathbf{z}$, could be extracted from the network as shown in Figure 1(b). The first embedding can be extracted only using speech input. The second embedding can be extracted using speech input with the special token as defined in our previous study [27] such as zero embedding ($\mathbf{e}_{\emptyset}$), random embedding ($\mathbf{e}_{\text{RAND}}$) and mean face embedding ($\mathbf{e}_{\text{MEAN}}$). We generated $\mathbf{e}_{\text{RAND}}$ drawn from a standard normal distribution and $\mathbf{e}_{\text{MEAN}}$ by averaging all face embeddings in the Voxceleb 2 data. These trivial embeddings were originally defined to mimic the special situation when one modality is missing. However, we found that they still help to generate a reasonable multimodal embedding $\mathbf{z}$.

For comparison, we also fine-tuned the model using only speech by adding a layer on top of the speaker verification model as shown in Figure 1(c). We used contrastive loss to train and followed the same hard negative mining approach as used in [1]. We tested this approach using telephone speech from the SRE04-10 data sets.

4.3. Experimental results

Table 3 (a) and (b) shows the experimental results based on the TDNN and E-TDNN x-vector models, respectively. We observed that the fine-tuning approach using only speech did not produce a substantial gain compared to the baseline x-vector system. This fine-tuning approach is usually applied to an end-to-end system that uses the cosine similarity measurement as a scoring backend between the two embeddings [1, 2]. We speculate that this approach shows a benefit in limited circumstances and has no advantage for systems that using PLDA as a back-end, so many successful studies related to SRE18 do not use this type of fine-tuning method. However, the proposed multimodal fine-tuning approach showed a significant improvement in the EER and C_{min} measurements. We observed that the proposed approach shows a consistent performance improvement on the two different x-vector systems based on the TDNN and E-TDNN. While the proposed approach can generate several variants of the embedding such as $\tilde{\mathbf{e}}_s$ and $\mathbf{z}$, it was observed that the $\tilde{\mathbf{e}}_s$ shows the best performance among them. Based on this observation, we verified that the fusion network could associate the multisensory data and make the unimodal system network perform better on the verification task.

[1] `https://github.com/davidsandberg/facenet`
We used this reproduced open model, which has been improved by the maintainers with several modifications. The modifications include a dimension change of the last layer from 128-D to 512-D. We use the last 512-D `FC7` layer activation of this FaceNet version as the face embedding.

[2] `http://dlib.net`

System	Development		Evaluation	
	EER(%)	C_{min}	EER(%)	C_{min}
x-vector (TDNN, $\mathbf{e}_s$)	6.64	0.453	7.44	0.513
Fine-tuning (SRE-Telephone)	6.52	0.431	7.26	0.502
Multimodal fine-tuning ($\tilde{\mathbf{e}}_s$)	**6.18**	**0.403**	**6.97**	**0.489**
Multimodal fine-tuning ($\mathbf{z}$ with $\mathbf{e}_{\emptyset}$)	6.35	0.418	7.13	0.500
Multimodal fine-tuning ($\mathbf{z}$ with $\mathbf{e}_{\text{RAND}}$)	6.40	0.425	7.21	0.503
Multimodal fine-tuning ($\mathbf{z}$ with $\mathbf{e}_{\text{MEAN}}$)	6.48	0.430	7.17	0.499

(a) *Based on the TDNN x-vector*

System	Development		Evaluation	
	EER(%)	C_{min}	EER(%)	C_{min}
x-vector (E-TDNN, $\mathbf{e}_s$)	5.80	0.377	6.57	0.470
Fine-tuning (SRE-Telephone)	5.81	0.375	6.59	0.471
Multimodal fine-tuning ($\tilde{\mathbf{e}}_s$)	**5.58**	**0.347**	**6.35**	**0.430**
Multimodal fine-tuning ($\mathbf{z}$ with $\mathbf{e}_{\emptyset}$)	5.69	0.359	6.52	0.444
Multimodal fine-tuning ($\mathbf{z}$ with $\mathbf{e}_{\text{RAND}}$)	5.78	0.351	6.43	0.441
Multimodal fine-tuning ($\mathbf{z}$ with $\mathbf{e}_{\text{MEAN}}$)	5.84	0.373	6.60	0.448

(b) *Based on the E-TDNN x-vector*

Table 3: *Performance on SRE18 CMN2 condition*

One limitation of this study is that we did not check to see if the face recognition model could also derive a benefit from the speaker verification model. In general, face recognition gives a more robust embedding than the speaker verification model, so we can expect the gradient by face in Equation (2) could act as the momentum of the stochastic gradient descent optimization. However, it is doubtful that the speaker verification model could also influence the face recognition model positively since the speaker embedding is not that robust compared to face embedding.

Another limitation of this study is that we did not evaluate the proposed approach on various channel conditions other than the telephony recording channel. This limitation originally is due to the lack of data to researchers for various evaluations. Note that even SRE data is only accessible by participants who have complete the evaluation event.

5. Conclusion

Motivated from a recent neuroscience study about multimodal association in humans, we proposed a multimodal fine-tuning approach for a speaker verification system inspired by human perception. The proposed approach mainly focused on the extraction of a robust and channel-invariant speaker embedding by preventing training dataset channel overfitting. Through an appropriate experimental design, we observed the multimodal association of ecologically valid and sensory redundant stimulus pairs can affect the unimodal perception network, i.e. speaker verification model. The proposed approach is based on fine-tuning of the pre-trained model, so it can be applied to many other methods previously introduced in other studies. Aside from the robust performance on the SRE18 benchmark, it can be potentially applied to many different domains since the proposed approach does not rely on the target domain channel characteristic. We believe further investigation is warranted for this approach since the multimodal association is not an explicit association between speech and visual features such as lip motion, but an implicit association between utterance level speech and face image. It would be also interesting to investigate the effect of speech on the face recognition.

6. References

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

[2] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep Speaker Recognition," in *Interspeech*, 2018, pp. 1086–1090.

[3] K. von Kriegstein and A.-L. Giraud, "Implicit multisensory associations influence voice recognition," *PLOS Biology*, vol. 4, no. 10, pp. 1–12, 09 2006.

[4] O. Glembek, J. Ma, P. Matejka, B. Zhang, O. Plchot, L. Burget, and S. Matsoukas, "Domain adaptation via within-class covariance correction in i-vector based speaker recognition systems," in *IEEE ICASSP*, 2014, pp. 4060–4064.

[5] H. Aronowitz, "Inter dataset variability compensation for speaker recognition," in *IEEE ICASSP*, 2014, pp. 4002–4006.

[6] ——, "Compensating Inter-Dataset Variability in PLDA Hyper-Parameters for Robust Speaker Recognition," in *Proceedings of Odyssey - The Speaker and Language Recognition Workshop*, 2014, pp. 280–286.

[7] D. Garcia-Romero and A. McCree, "Supervised domain adaptation for I-vector based speaker recognition," in *IEEE ICASSP*, 2014, pp. 4047–4051.

[8] J. Villalba and E. Lleida, "Unsupervised Adaptation of PLDA by Using Variational Bayes Methods," in *IEEE ICASSP*, 2014, pp. 744–748.

[9] D. Garcia-Romero, A. McCree, S. Shum, N. Brummer, and C. Vaquero, "Unsupervised Domain Adaptation for I-Vector Speaker Recognition," in *Proceedings of Odyssey - The Speaker and Language Recognition Workshop*, 2014, pp. 260–264.

[10] S. Shum, D. a. Reynolds, D. Garcia-Romero, and A. McCree, "Unsupervised Clustering Approaches for Domain Adaptation in Speaker Recognition Systems," in *Proceedings of Odyssey - The Speaker and Language Recognition Workshop*, 2014, pp. 265–272.

[11] M. H. Rahman, A. Kanagasundaram, D. Dean, and S. Sridharan, "Dataset-invariant covariance normalization for out-domain PLDA speaker verification," in *Interspeech*, 2015, pp. 1017–1021.

[12] A. Kanagasundaram, D. Dean, and S. Sridharan, "Improving out-domain PLDA speaker verification using unsupervised inter-dataset variability compensation approach," in *IEEE ICASSP*, 2015, pp. 4654–4658.

[13] S. Shon, S. Mun, W. Kim, and H. Ko, "Autoencoder based Domain Adaptation for Speaker Recognition under Insufficient Channel Information," in *Interspeech*, 2017, pp. 1014–1018.

[14] J. Zhou, T. Jiang, L. Li, Q. Hong, Z. Wang, and B. Xia, "Training multi-task adversarial network for extracting noise-robust speaker embedding," in *IEEE ICASSP*, 2019, pp. 6196–6200.

[15] Z. Chen, S. Wang, Y. Qian, and K. Yu, "Channel invariant speaker embedding learning with joint multi-task and adversarial training," in *IEEE ICASSP*, 2020, pp. 6574–6578.

[16] D. Cai, W. Cai, and M. Li, "Within-sample variability-invariant loss for robust speaker recognition under noisy environments," *arXiv preprint arXiv:2002.00924*, 2020.

[17] Z. Meng, Y. Zhao, J. Li, and Y. Gong, "Adversarial speaker verification," in *IEEE ICASSP*, 2019, pp. 6216–6220.

[18] S. Wang, J. Rohdin, L. Burget, O. Plchot, Y. Qian, and K. Yu, "On the usage of phonetic information for text-independent speaker embedding extraction," in *Interspeech 2019*, 2019, pp. 1148–1152.

[19] G. Bhattacharya, J. Monteiro, J. Alam, and P. Kenny, "Generative adversarial speaker embedding networks for domain robust end-to-end speaker verification," in *IEEE ICASSP*, 2019, pp. 6226–6230.

[20] S. Shon, H. Tang, and J. Glass, "Voiceid loss: Speech enhancement for speaker verification," *arXiv preprint arXiv:1904.03601*, 2019.

[21] C. Luu, P. Bell, and S. Renals, "Channel adversarial training for speaker verification and diarization," *arXiv preprint arXiv:1910.11643*, 2019.

[22] J. S. Chung, J. Huh, and S. Mun, "Delving into voxceleb: environment invariant speaker recognition," *arXiv preprint arXiv:1910.11238*, 2019.

[23] G. Bhattacharya, J. Alam, and P. Kenny, "Deep speaker recognition: Modular or monolithic?" in *Interspeech*, 2019, pp. 1143–1147.

[24] S. Yun, J. Cho, J. Eum, W. Chang, and K. Hwang, "An end-to-end text-independent speaker verification framework with a keyword adversarial network," *arXiv preprint arXiv:1908.02612*, 2019.

[25] K. Zhou, Q. Yang, X. Sun, and S. Liu, "A deep speaker embedding transfer method for speaker verification," in *The International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery*, 2019, pp. 369–376.

[26] L. Wang, Y. Wang, and M. J. Gales, "Non-native speaker verification for spoken language assessment," *arXiv preprint arXiv:1909.13695*, 2019.

[27] S. Shon, T.-H. Oh, and J. Glass, "Noise-tolerant audio-visual online person verification using an attention-based neural network fusion," in *IEEE ICASPP*, 2019, pp. 3995–3999.

[28] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *International Conference on Machine Learning*, 2011.

[29] D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, and Y. Carmiel, "Deep Neural Network Embeddings for Text-Independent Speaker Verification," in *Interspeech*, 2017, pp. 999–1003.

[30] D. Snyder, G. Chen, and D. Povey, "MUSAN : A Music , Speech , and Noise Corpus," *ArXiv e-prints arXiv:1510.08484*, 2015.

[31] J. Villalba, N. Chen, D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, J. Borgstrom, F. Richardson, S. Shon, F. Grondin, R. Dehak, L. P. G. Perera, D. Povey, P. Torres-Carrasquillo, S. Khudanpur, and N. Dehak, "State-of-the-art Speaker Recognition for Telephone and Video Speech: the JHU-MIT Submission for NIST SRE18," in *Interspeech*, 2019.

[32] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011.

[33] C. Zhang, F. Bahmaninezhad, S. Ranjan, H. Dubey, W. Xia, and J. H. Hansen, "Utd-crss systems for 2018 nist speaker recognition evaluation," in *IEEE ICASSP*, 2019, pp. 5776–5780.

[34] I. Viñals, D. Ribas, V. Mingote, J. Llombart, P. Gimeno, A. Miguel, A. Ortega, and E. Lleida, "Phonetically-aware embeddings, wide residual networks with time-delay neural networks and self attention models for the 2018 nist speaker recognition evaluation," in *Interspeech*, 2019, pp. 4310–4314.

[35] D. Cai, W. Cai, and M. Li, "The dku-smiip system for nist 2018 speaker recognition evaluation," *arXiv preprint arXiv:1907.02191*, 2019.

[36] K. A. Lee, H. Yamamoto, K. Okabe, Q. Wang, L. Guo, T. Koshinaka, J. Zhang, and K. Shinoda, "The NEC-TT 2018 Speaker Verification System," in *Interspeech*, 2019, pp. 4355–4359.

[37] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet : A Unified Embedding for Face Recognition and Clustering," in *Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 815–823.