Improvement of Text-Independent Speaker Verification Using Gender-like Feature

Pornprom Kiawjak*, Somkiat Wangsiripitak‡(✉) and Kitsuchart Pasupa†
Faculty of Information Technology
King Mongkut's Institute of Technology Ladkrabang
Bangkok 10520, Thailand
{59070113*, somkiat‡, kitsuchart†}@it.kmitl.ac.th

***Abstract*—Text-independent speaker verification is a task of verifying a speaker identity from a characteristic of voice. We proposed the combined deep Convolutional Neural Network (CNN) consisting of (i) the first CNN trained to achieve gender classification which is then used to create a gender-like embedding and (ii) the last CNN trained with one additional input, the gender-like feature (embedding) from the first, to classify each speaker. The classification layer of the last CNN is removed to allow the remaining combined deep CNN for one-shot learning and verification of unobserved speaker. Our proposed CNN could obtain better results compared to VGGVox (ResNet-50) by 0.40% of Equal Error Rate (EER) on average. Additionally, we investigated results based on the scenario that the gender is known; the evaluation was performed only on utterance pairs that comply with the scenario. The EER rate of such case that only gender of claimed identity is known is 0.52% lower than that of VGGVox (ResNet-50) on average of two genders. In a more specific situation that the gender of person making a claim is also known, two dedicated networks were retrained for female and male, in addition to our first network which was trained for both. It is interesting that, when compared to the first network, the female network achieved less EER on female-female verification, while the network dedicated for male performed worse. Nevertheless, our two dedicated networks outperformed VGGVox (ResNet-50) by 0.88% of EER.**

Keywords*—speaker verification; text-independent; gender-like feature; combined deep convolutional neural network (CNN)

I. Introduction

Speaker verification (SV) is a task which verifies whether a human voice is uttered from the claimed speaker. The concept of SV is a subset of speaker recognition (SR). It can be split into two modes: text-dependent and text-independent. In the scenario of text-dependence, a word, a phrase, or a sentence is designated as a target together with characteristics of a speaker's voice; the pre-defined text is used as a password constraint. On the other hand, the text-independent approach does not require the knowledge about what the utterance is. It uses only voice characteristics to distinguish between target speakers and imposters. Therefore, the text-independent SV is more challenging than the text-dependent SV, especially when it is used for authentication. In this work, we investigate on text-independent SV in particular.

Successful conventional methods for SV often apply unsupervised generative model or statistical model or both, such as the Gaussian Mixture Model-Universal Background Model (GMM-UBM) [1]. Some improved GMM with SVM classifier, e.g., [2]. All of traditional text-independent speaker verification system was explained well in [3]. Later, I-vector methods developed from GMM-UBM have been proposed (e.g. [4]) and then further extended by PLDA backend [5], [6] which helps increase distinctiveness of each speaker by assign more weight to features that more contributed—however the results may come from one of local minima because the distinctive features are not extracted/modeled from the beginning.

Recently, many deep learning based methods have been proposed in order to automatically construct a feature vector for samples of one class to be different from others, by learning from a large training data, which show promising results in many fields [7], [8], [9]. To avoid retraining of the neural network when there is a new class, d-vector based approaches [10], [11], [12] have been invented which allow a learning of new speaker with a few samples (e.g. utterances for SV). Some research groups utilized 3D-CNN (Convolutional Neural Network) architecture for text-independent SV which enables simultaneous training of many samples from one class. For example, Torfi *et al.* used such 3D-CNN architecture and claimed that it is more robust to within-speaker variation [13]. However, this method still needs more than one sample (speaker voice) to create a speaker model—it is unfavorable and impractical in some situations. Chung *et al.* proposed one of state-of-the-art method; They created VGGVox framework that could change CNN architecture and evaluated them on text-independent SV task. One-shot learning (one-sample learning) of newly observed person was used to create speaker model; it was used for speaker verification later [14], [15]. The best model of VGGVox is trained on one of well-known CNN architecture ResNet50 and evaluated on Minimum Detection Cost (minDCF) and Equal Error Rate (EER)—we used this as our baseline.

This paper proposes a novel approach for text-independent SV. Instead of data augmentation, voting, refining on architecture or time pooling layer, and/or selecting loss function that gives the best result for each architecture, we use gender-like features to help increase the accuracy of SV. We also investigate our approach in case we know speaker's gender. As SV for same gender tends to be worse (shown in our result), some modification (suggestion on how to utilize the proposed method in practice) is presented; it further increases the accuracy of SV.

II. Text-independent Speaker Verification Using Gender-like Feature

Several methods have been proposed for text-independent SV. Those SV methods are supposed to be used in many ways—one of them is for user authentication, e.g., when a bank clerk needs to verify whether the customer voice on a phone call is actually of the person that he/she is claiming to be. In such scenario, EER may be used to evaluate each method performance by applying SV on many utterance pairs in a data set. Several data sets have been prepared by many research groups for this purpose; however, in the scenario that a gender of person being claimed is known (e.g. when a bank clerk has information of all customers on database server, including a gender of customer being claimed), each utterance pair in the test data set should contain only pair of voices from same gender—it is unusual to verify if a woman voice is actually of the man being claimed.

Our study, as explained in the next subsection, shows that text-independent SV is more difficult when two utterances being matched are of the same gender—its error much more than that of different gender. We therefore proposed a gender-like feature which is used as a hard-constrained feature in addition to voice characteristics (features) extracted by deep learning network. Details of our approach are described subsequently.

A. Accuracy Drop in SV of Same Gender

VGGVox, one of state-of-the-art method, reported the results of SV on three different test sets in terms of EER; those are duplicated and shown in the first row of Table V. As shown in the table, the result of SV on VoxCeleb1-H test set is inferior to (EER is larger than) those of VoxCeleb1 and VoxCeleb1-E test sets. An obvious difference of these data sets is that: VoxCeleb1-H test set contains only utterance pair of the same gender and nationality; while those pairs in the other two test sets are arbitrary. According to these results, SV for two voice clips of the same gender is likely to be more error-prone than that of different gender—this motivates us to use gender information in learning and verification of speaker with the aim of improving the performance of SV for two voices of same gender.

B. Gender-like feature acquisition

In order to obtain a gender-like feature, we train a CNN (ResNet-50 is used in our experiment) with voice spectrogram for gender classification (two gender: female and male) as shown in Fig. 1. Each output from the CNN is normalized by sigmoid function; all results are passed into a loss function to measure binary cross entropy of the network during the training phase. The classification layer (fully connected layer) is then omitted, allowing the remaining to be used for extraction of gender-like feature (embedding) from voice (shown as a shaded box in Fig. 2).

C. Deep CNN training with gender-like feature

Traditional deep CNN for text-independent speaker verification consists of convolution, pooling, and fully connected layers (bottom half of Fig. 2)—an input of the network is voice spectrogram; the output is trained with a loss function. We emphasize the gender information of the voice by cascading the gender-like feature, which is extracted by the network in Fig. 1 (shown as shaded box in Fig. 2), and the traditional feature vector; finally, the combined feature vector is fed into the fully connected layer. To keep the sub-network for gender-like feature extraction (the shaded area in Fig. 2) unchanged because it has been trained already, the whole network is trained and fine-tuned (using Triplet loss) with this sub-network frozen. The feature vector that contains the distinct part for gender-like property hopefully helps increase the distance between feature vectors, each of which represents utterance of speaker (class) from the same gender; performance of speaker verification of two utterances of the same gender expectedly rises as the result.

D. Speaker verification

After training phase, we generalize our network for voices of new persons who have never been trained, by truncating all layers in the classification layers and beyond (in Fig. 2) except the first layer (512 nodes). The remaining network will generate 512-D vector; it is used for user (voice) enrollment and then for verification of person. The enrollment phase will register each speaker voice—here we use one-shot learning approach, i.e. only one utterance for each person is used for person enrollment. The verification phase is for verifying whether or not each new voice is of the enrolled person. The process starts from generating 512-D vector using the same network. That feature vector is then compared to the vector of enrolled person being claimed using cosine similarity; high score means two voices have more likeliness to be of the same person.

III. Experiments and results

Our proposed method is evaluated using EER metric. We assume that the method will be used for authentication of a customer that is talking to a staff either in person or on a phone—the staff may be a human or a bot. In such use case scenarios, the gender of the person being claimed by that customer is known because all customers' information is kept in database; however the gender of the man/woman who is talking to the staff may be known (if the staff could determine his/her gender by appearance or voice) or maybe not (if the staff is the bot that has no such ability). By taking into account those scenarios, we created two evaluation sets for both scenarios. Details are explained in Subsection III-B.

A. Settings

Our network requires voice spectrograms as an input. The spectrogram of each voice is computed from raw audio (16kHz) in a sliding window manner—here, hamming window of width 25 ms (millisecond) and step 10 ms is used. It produces 512×300 voice spectrograms for 3-second utterance. Our training data was organized to many batches; one batch has 64 utterances (each utterance is cropped from voice clip and has 3 seconds long). We trained them parallelly on two RTX 2080 TI GPUs for 30 epochs (or until validation loss converged). Stochastic Gradient Descent (SGD) with momentum equals to 0.9, weight decay of 5×10^{-4}, and learning rate with exponential decay (initialized to 10^{-2}) are employed in the experiments.

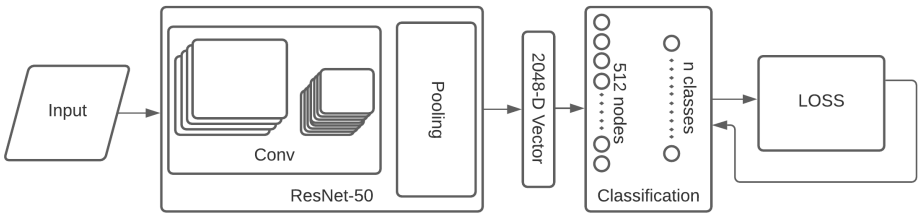

Fig. 1. Network for gender-like feature acquisition (training phase)

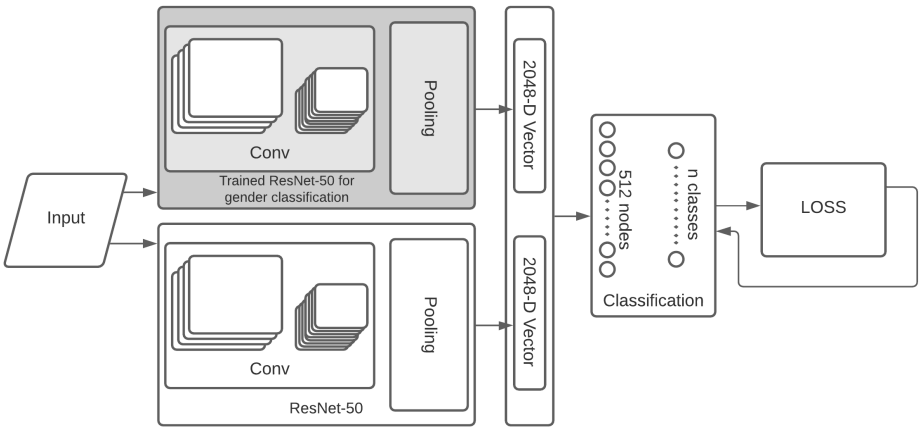

Fig. 2. Deep CNN with focused gender-like feature for text-independent speaker verification (training phase)

B. Dataset

Two publicly available datasets, VoxCeleb1 [14] and VoxCeleb2 [15], are used in the experiments. All utterances were retrieved from YouTube videos without post-processing. Therefore, some voice clips may be contaminated by background noises.

We used VoxCeleb2 (Dev), which has 5,994 classes (2,312 females and 3,682 males), in training and fine-tuning our network—there are 694,977 male utterances and 397,032 female utterances, as shown in Table I. 1,251 classes (2,312 females and 3,682 males) from VoxCeleb1 were used for performance evaluation—classes (persons) in VoxCeleb1 are all different from classes (persons) in VoxCeleb2 (Dev). Our experiments used three test sets—'original VoxCeleb1', 'VoxCeleb1-E' and 'VoxCeleb1-H'—from VoxCeleb1 (same as [15]) which contains many utterance pairs as detailed in Table II.

To evaluate the performance of our method under the scenarios explained above, we created two groups of utterance pairs, as shown in Table III and IV, which are rearranged from three test sets in Table II.

The first evaluation data group (Table III) assumes the scenario when we know the gender of person being claimed but not of the person making a claim. If someone claims to be some woman registered in the system, the first two rows of Table III will be used for evaluation (case A1). If the man is the target of being claimed, the last two rows is used (case A2)

TABLE I. STATISTICS OF 'VOXCELEB2 (DEV)' DATASET WHICH IS USED FOR TRAINING. EACH SAMPLE IS AN UTTERANCE.

	Female	Male	Total
No. of utterances	397,032	694,977	1,092,009
No. of classes	2,312	3,682	5,994

TABLE II. STATISTICS OF 'VOXCELEB1' TEST SET. EACH SAMPLE IS AN UTTERANCE PAIR.

Gender in a pair	Original VoxCeleb1		VoxCeleb1-E		VoxCeleb1-H	
	Class in a pair					
	different	same	different	same	different	same
Same (Female)	1,524	5,512	50,685	121,587	113,598	113,602
Same (Male)	9,228	13,348	98,856	169,156	162,668	162,668
Different	8,108	-	141,196	-	-	-
Sub-total	18,860	18,860	290,737	290,743	276,226	276,270
Total	37,720		581,480		552,536	

TABLE III. STATISTICS OF 'VOXCELEB1' TEST SET WHICH IS REARRANGED ACCORDING TO THE TEST SCENARIO THAT THE GENDER OF PERSON BEING CLAIMED IS KNOWN

Case	Gender in a pair	Original VoxCeleb1	VoxCeleb1-E	VoxCeleb1-H
A1	Same (Female)	7,036	172,272	227,200
	Different	8,108	141,146	-
A2	Same (Male)	22,576	268,012	325,336
	Different	8,108	141,146	-

TABLE IV. STATISTICS OF REORGANIZED 'VOXCELEB1' TEST SET ON THE CONDITION THAT THE GENDERS OF PERSONS MAKING A CLAIM AND BEING CLAIMED ARE KNOWN. ONLY PAIRS OF SAME GENDER UTTERANCES ARE REQUIRED IN THIS TEST SCENARIO.

Case	Gender in a pair	Original VoxCeleb1	VoxCeleb1-E	VoxCeleb1-H
B1	Same (Female)	7,036	172,272	227,200
B2	Same (Male)	22,576	268,012	325,336

The other group (Table IV) assumes that the gender of person making a claim and gender of user account being claimed are known; this group therefore contains only those utterance pairs of which each comprises of two voices of same gender—we defined female-female pairs as case B1 and male-male pairs as case B2 (no need to compare two utterances of different genders).

C. Training

The CNN for gender-like feature extraction in Fig 1 was trained using utterances from VoxCeleb2 (Dev) (Table I). This network—without classification layer—was then used for obtaining gender-like feature, which was later integrated into the SV network (Fig. 2). We performed three experiments on the SV network as to three scenarios of usage: when no gender prior is known (experiment I); when the gender of user account being claimed is known (experiment II); and when the genders of user account being claimed and person making a claim are known (experiment III). In order to serve those scenarios, the SV network was trained in two ways: (i) one SV network—later called 'general SV-CNN'—was trained without gender info, i.e. using the whole utterances from VoxCeleb2 (dev) (the last column in Table I); and (ii) two instances of dedicated SV network—later called 'male SV-CNN' and 'female SV-CNN'—were trained separately, each used utterances of only one gender (male or female: the first two columns of Table I). Note that utterances of each person in VoxCeleb2 (dev) comes from several recording scenarios. We randomly selected one recording for each person and used all utterances from that recording for validation; the remaining were used in the training. The network was trained and validated until it converged; it is then fine-tuned using triplet loss. The computational complexity of the training in one epoch, in terms of time, is 7.5, 2.0, and 4.5 hours for the 'general SV-CNN', 'female SV-CNN', and 'male SV-CNNN' respectively.

D. Evaluation

The network which was trained and fine-tuned will be evaluated using VoxCeleb1 test set that contains many utterance pairs—two voices in the pair may be of the same or different person. One utterance in the pair is for one-shot learning of user voice and the other for verification. Since the persons in VoxCeleb1 are totally different from those of VoxCeleb2 (dev), generalization of the trained network will be proved with low error rate. Note that the performance of our approach is measured in terms of EER. The computational complexity of the one-shot learning in registration of user voice or creating an embedding vector of incoming user in authorization, in terms of time, is 2.43 seconds (average of 27 arbitrary-lengthy utterances). Details of experiments on three usage scenarios are explained next.

TABLE V. PERFORMANCE (EER%) OF VGGVOX AND GENERAL SV-CNN (OUR METHOD)

Method	Original VoxCeleb1	VoxCeleb1-E	VoxCeleb1-H
VGGVox	**3.95**	4.42	7.33
General SV-CNN (our method)	4.08	**3.86**	**6.54**

TABLE VI. NUMBERS OF ERRORS (TOP VALUE) AND ERROR RATES (BELOW VALUE IN PARENTHESIS) OF GENERAL SV-CNN (OUR METHOD)

Gender in a pair	Original VoxCeleb1		VoxCeleb1-E		VoxCeleb1-H	
	Class in a pair					
	different	same	different	same	different	same
Same (Female)	150 (9.84%)	129 (2.34%)	4,483 (8.84%)	4,500 (3.70%)	9,131 (8.03%)	7,071 (6.22%)
Same (Male)	616 (6.67%)	641 (4.80%)	6,358 (6.43%)	6,738 (3.98%)	8,953 (5.50%)	11,014 (6.77%)
Different	4 (0.05%)	-	395 (0.28%)	-	-	-
Sub-total	770 (4.08%)	770 (4.08%)	11,236 (3.86%)	11,238 (3.86%)	272,266 (6.54%)	272,270 (6.54%)

E. Experiment I: (general scenario)

Our general SV-CNN, which was fine-tuned already, was evaluated on three test sets. The results are shown in Table V. The results of VGGVox method [15] are also shown here for comparison. Although our general SV-CNN is slightly inferior to VGGVox on original VoxCeleb1 test set, it is better (EER is less). The number of utterance pairs in original-VoxCeleb1 (37,720 pairs) is far less than those of VoxCeleb1-E (581,480 pairs) and VoxCeleb1-H (552,536 pairs); therefore, our general SV-CNN is generalized better than VGGVox by inference. It is worth noting that our 'general SV-CNN' obtained 0.40% less EER than VGGVox on average.

Table VI shows more details of our performances; numbers of errors and error rates (in percentage) of our general SV-CNN on three subgroups of those test sets are shown. According to the results, verification of two utterances of the same gender but different person has more error than those of the same person—i.e. false-positive rate (FPR) is larger than the false-negative rate (FNR) when trying to verify two voices of the same gender. This does not align with the overall EER in Table V (the overall EER is measured on all cases; its value is repeated in the 'Sub-total' row of Table VI). It means unauthorized persons are granted access more than expected. It is not desirable, especially when the gender is known and most verifications are performed on two utterances of the same gender. We therefore proposed the training method that focuses on this scenario; experimental results are shown in next subsections.

TABLE VII. PERFORMANCE (EER%) OF OUR GENERAL SV-CNN (OUR METHOD) WHEN ONLY GENDER OF USER BEING CLAIMED IS KNOWN.

Case	Gender of two utterances in the pair	Method	OriginalVoxCeleb1	VoxCeleb1-E	VoxCeleb1-H
A1	Female–Unisex	General SV-CNN	1.87	3.05	7.03
A2	Male–Unisex	General SV-CNN	4.13	3.37	6.17
Weighted average		General SV-CNN	3.38	3.23	6.52

TABLE VIII. PERFORMANCE (EER%) COMPARISON OF OUR SV-CNNS (GENERAL AND DEDICATED NETWORKS) WHEN THE GENDER OF USER BEING CLAIMED AND THAT OF PERSON MAKING A CLAIM ARE KNOWN.

Case	Gender of two utterances in the pair	Our method (SV-CNN)	OriginalVoxCeleb1	VoxCeleb1-E	VoxCeleb1-H
B1	Female–Female	General	5.04	5.71	7.03
		Dedicated (female)	4.33	5.66	6.34
B2	Male–Male	General	5.60	4.99	6.17
		Dedicated (male)	6.29	5.43	6.53
Weighted average		General	5.46	5.27	6.52
		Dedicated (female & male)	5.82	5.51	6.45

F. Experiment II: (only gender of claimed user is known)

In this experiment, we assumed that someone (male or female) tries to use our unattended system by claiming that he or she is one of our customer. Our system knows the gender of our customer (by checking the information in database) but not of the coming user. Two separate scenarios could happens: one is when the gender of user being claimed is female (case A1) and the other is when the account being claimed is of male (case A2). In such cases, we can use our 'general SV-CNN' (the same network used in experiment I) with two decision thresholds—each is dedicated to each case (this is possible because the system knows the gender of user being claimed). Each threshold is obtained during the evaluation of our general SV-CNN on utterance pairs that satisfy the gender constraint of corresponding case, i.e. finding the best threshold that give the lowest EER. The results, in terms of EER, on three VoxCeleb1 test sets are shown in Table VII. Weighted average EERs are also listed in the last row. Note that the EER of VoxCeleb1-H is approximately twice of the other two—this test set contains only utterance pairs of same gender, which is more difficult to correctly verify the speaker.

As shown in Table VII and V, the weighted average EER of general SV-CNN with two thresholds (the last row of Table VII) is less than the EER of the same network with one threshold (the last row of Table V). Provided that the gender of user being claimed is known, our general SV-CNN with two dedicated thresholds (one for matching the incoming user with female customer in the database; the other with male customer) is therefore preferable to general SV-CNN with one threshold. It is noteworthy that our twin-threshold 'general SV-CNN' obtained 0.52% less EER than VGGVox on average.

G. Experiment III: (genders of claimed user and claiming person are known)

Assuming that a customer comes to a service counter to make a transaction. He (or she) declares himself (or herself) as one of the customer. In such case, a counter staff knows the gender of that customer and also that of the one he (or she) is claiming to be. This experiment assumed such situation. Therefore the performance of speaker verification should evaluated only on a pair of utterances that belongs to a same gender—matching of one gender's voice to the other gender's is an invalid case and will never happen. The test sets that are reorganized according to the above-mentioned situation are shown in Table IV: case B1 contains only pairs of two female voices while case B2 consists of male utterance pairs.

As did in experiment II (see Subsection III-F), we used the 'general SV-CNN' with two decision thresholds, but here it was evaluated on speaker verification of only same gender. We also used our two dedicated SV-CNNs—one is 'female SV-CNN' and the other is 'male SV-CNN' (they are the retrained networks which were explained in Subsection III-C)—for speaker verification of the same test sets. Experimental results are shown in Table VIII. The 'female SV-CNN' is superior to the 'general SV-CNN' when matching two utterances of female; this means the dedicated network trained specifically for female-female speaker verification can increase the ability of classification. On the other hand, it is interesting that the 'dedicated SV-CNN' for male has higher EER than the 'general SV-CNN'—the reason is needed to be investigated. According to these outcomes, one may utilize 'female SV-CNN' for female–female verification, and 'general SV-CNN' for male-male.

It is worth mentioning that, here, the result on OriginalVoxCeleb1 and VoxCeleb1-E test sets cannot be compared to those of experiments I and II, because utterance pairs (of the same gender) used in this experiment are only subset of the whole test sets used in those experiments. However, we can compare the results on VoxCeleb1-H since this test set contains only voice pairs of the same gender. On that test set, our two dedicated SV-CNNs achieved EER of 6.45% on average—this is better than VGGVox and our 'twin-threshold general SV-CNN' which obtained EER of 7.33% and 6.52% respectively.

IV. CONCLUSIONS

In this paper, we introduced a combined CNN with a gender-like feature for text-independent speaker verification. We demonstrated our proposed method on the VoxCeleb1 test sets and compared the result to VGGVox. The generalization of our network to verification of utterance with one-shot learning is superior to VGGVox; this includes the special case when two voices being matched are of the same gender (VoxCeleb1-H test set). In addition, we evaluated our network in particular situations: when the gender of claimed user can be read from a database, and when the gender of customer can also determined (maybe by a service person). Based on this prior information, the proposed network could be tuned or retrained for matching of each specific gender; the evaluation on valid test cases corresponding to each situation showed more promising results. We have found another state-of-the-art architecture that improved time-aggregation [16] and had better results when compared to VGGVox; our preliminary analysis of error reveals that its improvement on VGGVox is partly different from ours, e.g. in terms of error mode. Further investigation on this is to be done in the future.

REFERENCES

[1] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, Jan. 2000.

[2] W. Campbell, D. Sturim, and D. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 308–311, May 2006.

[3] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. Petrovska-Delacrétaz, and D. A. Reynolds, "A Tutorial on Text-Independent Speaker Verification," *EURASIP Journal on Advances in Signal Processing*, vol. 2004, no. 4, p. 101962, 2004.

[4] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint Factor Analysis Versus Eigenchannels in Speaker Recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.

[5] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of I-vector Length Normalization in Speaker Recognition Systems," in *Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH 2011)*, 2011, pp. 249–252.

[6] O. Novotny, O. Plchot, O. Glembek, L. Burget, and P. Matejka, "Discriminatively Re-trained I-vector Extractor for Speaker Recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2019)*. Brighton, United Kingdom: IEEE, 2019, pp. 6031–6035.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[8] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015, pp. 815–823.

[9] X. Dong and J. Shen, "Triplet Loss in Siamese Network for Object Tracking," in *Proceedings of the European Conference on Computer Vision (ECCV 2018)*, ser. Lecture Notes in Computer Science, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., vol. 11217. Cham: Springer International Publishing, 2018, pp. 472–488.

[10] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2014)*. Florence, Italy: IEEE, 2014, pp. 4052–4056.

[11] Y.-h. Chen, I. Lopez-Moreno, T. N. Sainath, M. Visontai, R. Alvarez, and C. Parada, "Locally-Connected and Convolutional Neural Networks for Small Footprint Speaker Recognition," in *Proceedings of the 16th Annual Conference of the International Speech Communication Association (INTERSPEECH 2015)*, 2015, pp. 1136–1140.

[12] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*. Shanghai: IEEE, Mar. 2016, pp. 5115–5119.

[13] A. Torfi, J. Dawson, and N. M. Nasrabadi, "Text-Independent Speaker Verification Using 3D Convolutional Neural Networks," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME 2018)*, 2017.

[14] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A Large-Scale Speaker Identification Dataset," in *Proceedings of the 18th Annual Conference of the International Speech Communication Association (INTERSPEECH 2017)*. ISCA, 2017, pp. 2616–2620.

[15] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep Speaker Recognition," in *Proceedings of the 19th Annual Conference of the International Speech Communication Association (INTERSPEECH 2018)*. ISCA, Sep. 2018, pp. 1086–1090.

[16] W. Xie, A. Nagrani, J. S. Chung, and A. Zisserman, "Utterance-level Aggregation For Speaker Recognition In The Wild," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2019)*, May 2019.