Le comportement des systèmes de reconnaissance du locuteur de l'état de l'art face aux variabiliés acoustiques

Mohammad Mohammadamini, Driss Matrouf, Jean-François Bonatsre, Sandipana Dowerah, Romain Serizel, Denis Jouvet

Le comportement des systèmes de reconnaissance du locuteur de l'état de l'art face aux variabiliés acoustiques

Mohammad Mohammadamini[1] Driss Matrouf[1] Jean-François Bonatsre[1]
Sandipana Dowerah[2] Romain Serizel [2] Denis Jouvet [2]
(1) Laboratoire Informatique d'Avignon
(2) University of Lorraine, CNRS, Inria, Loria, F-54000, Nancy, France
mohammad.mohammadamini, driss.matrouf,
jean-francois.bonastre@univ-avignon.fr, sandipana.dowerah,
romain.serize@loria.fr, denis.jouvet@inria.fr

RÉSUMÉ

Dans cet article, nous étudions la résistance des systèmes de reconnaissance du locuteur de l'état de l'art face aux variabilités acoustiques, telles que le bruit additif et la réverbération. Deux systèmes seront comparés, le premier est fondé sur TDNN tandis que le second est fondé sur ResNet. Nous montrerons que globalement et sans utilisation de techniques de compensation le ResNet est plus robuste aux bruits que le TDNN. Les techniques de compensation dans le cas des TDNN sont efficaces lorsque les bruits sont ajoutés, alors qu'elles n'apportent aucun gain dans le cas du ResNet. En outre, les performances obtenues avec ResNet sans compensation restent significativement supérieures à celles d'un TDNN, même avec compensation. Les expériences sont réalisées en utilisant les corpus Fabiol et Voices. Des visualisations graphiques et des expériences complémentaires sont réalisées afin de donner des explications à ces différences de comportement face aux bruits du TDNN et du ResNet.

ABSTRACT

The behaviour of the state-of-the art speaker recognition systems against acoustic variabilities

In this paper we study the resistance of the state-of-the-art speaker recognition systems against acoustic variables such as additive noise and reverberation. Two systems will be compared : the first one is based in TDNN and the second one is based in ResNet. We show that generally and without using compensation techniques the ResNet is more robust against noise than TDNN. The compensation techniques are efficient for simulated noise with TDNN, while they don't bring gain with ResNet. Furthermore, the obtained performance using ResNet without compensation is superior to TDNN with compensation. The experiments are done using Fabiol and Voices corpus. Graphical visualisations and complementary experiments are done to explain these differences between TDNN and ResNet against noise.

MOTS-CLÉS : Reconnaissance de locuteur, ResNet, Bruit additif, Réverbération, Robustesse.

KEYWORDS: Speaker recognition, ResNet, Additive noise, Reverberation, Robustness.

1 Introduction

La reconnaissance du locuteur consiste à authentifier un locuteur à travers sa voix. Les systèmes de reconnaissance du locuteur de l'état de l'art utilisent des DNN (Deep Neural Network) pour extraire une représentation vectorielle de taille fixe à partir d'un signal de taille variable. Ces vecteurs appelés *Embeddings* représentent les locuteurs. Les distances entre ces *Embeddings* permettent de définir les proximités entre les locuteurs. L'estimation des *Embeddings* robustes aux bruits est une tâche essentielle dans les systèmes de reconnaissance du locuteur. Depuis l'émergence du système des x-vecteurs utilisant des TDNN (Snyder *et al.*, 2018) jusqu'à maintenant, plusieurs architectures de DNN pour l'estimation des *Embeddings* ont été proposées. Les systèmes TDNN (Snyder *et al.*, 2018), CNN (D. Cai & Li, 2018), ResNet et VGGVox (Arsha Nagrani, 2020) sont couramment utilisés.

La robustesse des systèmes de reconnaissance du locuteur (SR) basés sur les DNN en général et plus particulièrement leur robustesse face aux variabilités de l'environnement acoustique telles que le bruit additif, la réverbération et la distance du microphone les a rendus plus prometteurs. Plusieurs stratégies telles que l'augmentation des données (Mohammadamini, 2021b) et la compensation du bruit (Mohammadamini, 2020, 2021a) sont explorées pour rendre les systèmes basés sur TDNN plus robustes contre le bruit ambiant, la réverbération et d'autres variabilités.

Les recherches précédentes montrent la faiblesse des SR basés sur les TDNN contre les distorsions causées par le bruit et par la réverbération. Dans (Mohammadamini, 2020, 2021a), il est démontré qu'en présence de bruit et de la réverbération, l'utilisation d'une technique de compensation avant le *scoring* (Compensation avec un modèle statistique ou DAE (Denosing Auto-Encode)) peut grandement améliorer les performances, au point de s'approcher de celles obtenues avec des x-vecteurs propres (non bruités).

Dans cet article, nous commençons par une étude expérimentale de la robustesse des SRL utilisant des *Embeddings* issus de ResNet. Ensuite, nous étudions la possibilité de compenser les variabilités acoustiques au sein des *Embeddings* issus du ResNet. Dans ce contexte, nous comparons les *Embeddings* issus des TDNN (Snyder *et al.*, 2018) et ceux issus des ResNet. Nous allons montrer que le ResNet est plus robuste que le TDNN. Nous montrerons aussi que les techniques de compensation qui ont montré leur efficacité dans le cas du TDNN n'apportent aucun gain supplémentaire dans le cas du ResNet. Nous montrerons surtout que les performances du ResNet (sans compensation) restent meilleures que celles du TDNN même avec compensation.

Nos visualisations de type t-SNE (Laurens van der Maaten, 2008) des x-vecteurs, montrent que dans le cas du ResNet les x-vecteurs bruités et les x-vecteurs propres se situent dans le même région de l'espace, on peut même dire qu'ils sont entremêlés. Alors que dans le cas du TDNN, les ensembles de x-vecteurs se trouvent dans deux régions très distincts. Ce qui incite à penser que le ResNet est relativement robuste aux variations acoustiques, en comparaison avec le TDNN. Ceci est en conformité avec les résultats expérimentaux, en terme de performance, obtenus par les deux systèmes.

Dans la suite de cet article, dans la section 2, les travaux en lien avec notre problématique sont passés en revue. Dans la section 3, les systèmes étudiés dans ce travail sont décrits. Dans la section 4, nous décrivons les différentes configurations expérimentales utilisées. Dans la section 5, les résultats expérimentaux et les discussions sont exposés. Nous finissons par une conclusion dans la section 6.

2 Travaux connexes

La robustesse d'un système de reconnaissance du locuteur est traitée dans différentes parties du système, notamment au niveau du signal, au niveau de l'extraction des paramètres acoustiques, au niveau de l'extraction (et/ou la compensation) des *Embeddings*, au niveau du *scoring*. Dans cette section, les travaux connexes sont passés en revue. Les travaux examinés se répartissent en deux catégories principales : les techniques de compensation au niveau des *Embeddings* et l'extraction de *Embeddings* robustes aux bruits.

La première génération d'utilisation de DNN pour l'extraction de représentations vectorielles pour les locuteurs est fondée sur les TDNN (Snyder *et al.*, 2018). Différentes variantes de ce système ont été proposées : E-TDNN (D. Snyder, 2019), qui utilise un contexte temporel plus important et FTDNN (ref, 2018) qui à chaque couche du DNN prend en compte les informations provenant des couches précédentes, en imposant des contraintes d'orthogonalité sur les matrices de passage entre les couches. Le ResNet est une nouvelle architecture qui a rencontré un succès indéniable dans la modélisation vectorielle de locuteurs (K. He, 2016). Dans (Zhongxin Bai, 2021) plusieurs améliorations sont passés en revue ; certaines agissent sur les paramètres d'entrée, d'autres sur le critère d'optimisation (*loss function*) ou sur la modification de l'opération d'agglomération (*pooling*). Dans ces travaux, il n'y a pas de traitement ou de modélisation explicite du bruit ou de la réverbération : aucun modèle de bruit n'est utilisé.

Peu de travaux ont tenté de traiter le bruit d'une manière directe au niveau des *Embeddings*. Dans (Minh Pham & Whitehill, 2020),il est démontré que l'apprentissage d'*Embeddings* en utilisant des bruits spécifiques a très peu d'impact sur les performances du SRL. On montre que l'utilisation de données incluant le plus de variabilités acoustiques dans l'apprentissage conduit à de meilleurs résultats. Dans (Suwon Shon, 2019) une fonction de coût appelée VoiceID a été proposée, elle utilise la sortie du système d'extraction des x-vecteurs pour générer des coefficients appelés ratio de masque (*ratio mask*) permettant de réaliser un filtrage dans le domaine spectral.

En général, la plupart des systèmes tentent de produire une représentation (*Embedding*) la plus robuste possible des enregistrements vocaux, indépendamment de la qualité de l'environnement acoustique. Compenser directement le bruit, la réverbération ou d'autres variabilités au niveau des *Embeddings* est une autre approche qui pourraient réduire l'impact du bruit sur les systèmes de reconnaissance du locuteur. Dans (Mohammadamini, 2020) une compensation utilisant une approche bayesienne ou de DAE (*Denoising Auto-Encoder*) a été utilisée afin d'appliquer une transformation permettant d'obtenir une estimation d'un x-vecteur propre à partir d'une version bruitée de celui-ci. Dans (Mohammadamini, 2021b), il a été montré que les techniques d'augmentation des données (*data augmentation*) rendent les SRL plus robustes, mais que l'utilisation de techniques de débruitage (ou de compensation) apportent un gain significatif supplémentaire.

3 Les systèmes étudiés

Dans la section 3.1 l'architecture du ResNet est présentée. Dans la section 3.2 l'intégration du module de compensation des bruits au système de reconnaissance du locuteur est présentée.

3.1 Les architectures ResNet et TDNN

La reconnaissance du locuteur a connu ces dernières années des progrès notables grâce, en premier lieu,à l'introduction des réseaux de neurones (DNN) : un enregistrement est représenté par un vecteur appelé *embedding* qui contient essentiellement des informations concernant le locuteur (x-vecteur).

Les x-vecteurs du ResNet utilisé dans cet article est une variante du ResNet présenté dans (Zeinali, 2019). Le principe est le suivant : après avoir transformé le signal de parole en une suite de vecteurs acoustiques (dans notre cas, des bancs de filtres), le réseau de neurones prend comme entrée les vecteurs acoustiques et tente d'optimiser la fonction de coût qui n'est autre que la cross-entropie de la classification en locuteur. Autrement dit, les paramètres du réseau (DNN avant le *Embedding* et le classifieur) sont estimés de façon à minimiser le taux d'erreurs de classification des locuteurs étant donnés les signaux correspondants. Le ResNet est constitué principalement d'une suite de blocs (appelés blocs résiduels) dont chacun est constitué de deux convolutions séparées par une non-linéarité (ReLU) (He *et al.*, 2016). L'entrée d'un bloc est additionnée à sa sortie pour constituer l'entrée du bloc suivant. C'est grâce à cette dernière propriété que l'on arrive à augmenter la profondeur du réseau (encore et encore) tout en continuant à gagner en performances. Le RESNET est constitué (entre autres) de plusieurs dizaines de blocs résiduels, dans notre cas on en compte 34. Au niveau du *Embedding*, une opération de moyennage (mean-pooling) est appliquée. Plus de détails sont donnés dans (Brignatz V., 2021).

Dans nos expériences, nous avons utilisé l'architecture TDNN introduite dans (Snyder *et al.*, 2018). Le réseau TDNN composé de couches au niveau de la trame, au niveau de regroupement statistique (*statistical pooling*) et au niveau du segment. Les couches au niveau de la trame utilisent une architecture de temporisation qui tient compte du contexte temporel. La couche de regroupement statistique calcule la moyenne et l'écart type sur les trames de l'enregistrement vocal. Au niveau du segment, deux couches entièrement connectées sont ajoutées et une couche softmax est utilisée à la sortie pour la classification des locuteurs.

3.2 Module de compensation

Le module de compensation effectue une transformation entre les x-vecteurs bruités et les x-vecteurs propres. Ce faisant, le module de compensation essaie de supprimer l'impact du bruit dans les x-vecteurs. Dans ce travail, deux techniques sont testées pour la compensation : i-MAP et DAE. La technique statistique i-MAP qui est déjà utilisée pour les i-vecteur ou les x-vecteur (Waad Ben Kheder, 2017). L'approche fondée sur les DAE est décrite en détail dans (Mohammadamini, 2021b). En phase expérimentale, nous avons testé plusieurs variantes de cette architecture pour la compensation du bruit dans les x-vecteurs obtenus avec un système à base de ResNet. Ces détails seront discutés dans la section 5.2.

4 Configurations experimentales

4.1 Apprentissage des *Embeddings*

Les extracteurs des x-vecteurs sont entraînés sur le corpus Voxceleb. Afin d'augmenter la diversité des conditions acoustiques dans l'ensemble d'apprentissage, le corpus MUSAN a été utilisé pour l'augmentation des données (data augmentation) (D. Snyder, 2015). Aussi, un ensemble de RIR (Room Impulse Response) est utilisé pour la réverbération des données (Snyder *et al.*, 2018). Le système d'extraction des x-vecteurs TDNN utilise une paramétrisation de type MFCC (avec 50 coefficients et une durée de 25ms), et le ResNet utilise pour chaque trame 60 log Energies sorties de bancs de filtres (60 Log Filter Banks outputs).

Pour l'apprentissage du DNN, nous avons utilisé le corpus Voxceleb 1 et 2, contenant plus d'un million de sessions et plus d'un millier de locuteurs. Afin d'augmenter la robustesse du système, les données d'apprentissage ont été augmentées (Data Augmentation) en ajoutant d'une manière artificielle du bruit et de la réverbération aux enregistrements de Voxceleb 1 et 2. Au final environ 5 millions d'enregistrements ont été utilisés pour l'apprentissage.

4.2 Segments de test et Segments d'apprentissage des clients

Dans nos expériences, nous avons utilisé deux jeux de données. Le corpus Fabiol est utilisé pour évaluer la robustesse du système face au bruit et à la réverbération simulés. Dans le protocole Fabiol, nous avons 130 locuteurs clients dont 30 locuteurs utilisés pour le test. Le nombre de fichiers de test est de 6870. Un seul fichier par locuteur est utilisé pour l'apprentissage du modèle locuteur (Embedding). Le protocole Voices est utilisé pour évaluer la robustesse face au bruit et à la réverbération réels. L'ensemble de données Voices (Colleen Richey, 2018) comporte des parties d'entraînement et de test. La partie test a été créée à partir de 1320 fichiers propres provenant de Librispeech (100 locuteurs) et la partie train est enregistrée à partir de 2583 fichiers provenant de Librispeech (200 locuteurs). Nous avons utilisé 300 fichiers, chaque fichier appartenant à un locuteur pour l'enrôlement et 3603 fichiers restants ont été utilisé pour le test. Dans toutes les expériences utilisant le corpus Voices, le microphone éloigné (mic 05) et les pièces (room2, room3) avec plus de réverbération sont choisis. Les détails des protocoles sont présentés dans Table 1.

4.3 Back-end

Dans cet article, le système TDNN utilise un scoring avec PLDA. La PLDA est entraînée en utilisant 200k sessions extraites de Voxceleb. Le système ResNet utilise un scoring plus simple basé sur une simple distance cosinus.

5 Résultats et discussion

Dans cette section, une description des bases de données utilisées ainsi que les résultats expérimentaux dans différentes situations de bruits sont présentés. L'objectif est de montrer le comportement face aux bruits du TDNN en comparaison avec le ResNet.

TABLE 1 – Les protocoles d'évaluation.

Protocole	Nombre de segments de test	Nombre de clients	Trials
Fabiol	6870	130	893k
Voices	3603	300	1080k

TABLE 2 – Système de base.

Protocole	ResNet	TDNN
Fabiol	6.27	15.21
Voices	0.89	1.25

Baseline. Dans l'expérience de base, il n'y a ni bruit ni réverbération. Les résultats pour les protocoles Fabiol et Voices sont présentés dans le tableau 2.

Bruit additif. A notre connaissance, il n'existe pas de corpus avec seulement du bruit additif. C'est pour cette raison que nous testons les systèmes avec du bruit additif ajouté. Les expériences sont réalisées avec les jeux de données de bruit BBC (Noise, 2018) et Freesound. Dans trois expériences, différents SNR sont testés. Les résultats sont présentés dans le tableau 3.

Réverbération. Dans une autre expérience, nous avons exploré la robustesse des deux systèmes contre la réverbération. Dans ce cas, nous avons testé les systèmes avec une réverbération réelle et une réverbération simulée. Le protocole d'ajout de réverbération est décrit dans (Mohammadamini, 2021a). Pour la réverbération réelle, nous avons utilisé des enregistrements effectués dans "room2" et dans "room3" du corpus Voices sans bruit additif. Les résultats sont présentés dans le tableau 4.

Présence simultanée de bruit et de réverbération. Dans ce cas, les systèmes sont testés avec du bruit additif et de la réverbération réels et simulés. Les résultats sont présentés dans le tableau 5.

La comparaison des résultats de la Table 2 avec les tables Table 3, Table 4, and Table 5 nous constatons que les bruits (additif et réverbération) affectent les performances des deux systèmes (ResNet et TDNN). Mais ce que l'on remarque, en particulier, c'est que la robustesse aux variabilités acoustiques du ResNet est beaucoup plus importante que celle du TDNN. Par exemple, dans la pire situation (Fabiol [SNR 0-5]), le taux d'égales erreurs du ResNet est de 12.48% alors que celui du TDNN est de 21.47% (EN l'absence de bruits les taux sont respectivement 6.27% et 15.21%). Dans toutes les autres expériences, ResNet montre une grande robustesse contre le bruit et la réverbération. Par exemple, dans le protocole Voices sans bruit et sans réverbération, l'EER est de 0,89% mais en

TABLE 3 – Robustesse face au bruit additif dans Fabiol (EER)

SNR	BBC	BBC	Freesound	Fressound
-	ResNet	TDNN	ResNet	TDNN
0-5	8.31	18.15	8.28	17.83
5-10	7.43	16.58	7 :37	16.55
10-15	6.87	15.95	6.94	15.98

TABLE 4 – Robustesse face à la réverbération (EER)

Protocol	ResNet	TDNN
Fabiol	9.75	-
Voices room 2	1.24	2.53
Voices room 3	2.6	6.68

TABLE 5 – Résultats avec bruit additif et réverbération (EER)

Protocol	ResNet	TDNN
Fabiol [SNR 0-5]	12.48	21.47
Voices room 2	1.24	3.71
Voices room 3	2.6	6.69

présence de bruit sévère et de réverbération, il n'est que de 2,6%.

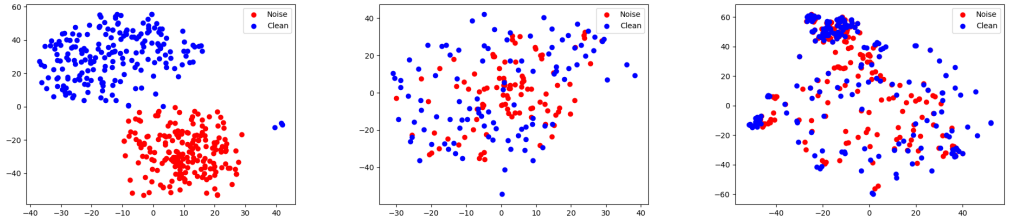
Afin de réduire l'effet des bruits sur les systèmes, nous effectuons une compensation du bruit au niveau des x-vecteurs. Les résultats après compensation du bruit pour les deux systèmes sont présentés dans le Table 6. Dans nos expériences, nous avons utilisé des paires de x-vecteurs (bruité, propres) pour entraîner le DAE (DNN) et l'i-MAP (Modèle statistique). Nous n'y avons observé aucun gain en termes d'EER dans cas du Resnet. Par exemple, en présence de bruit additif et de la réverbération dans le protocole Fabiol, l'EER passe de 12,48 avant compensation à 12,18 après compensation, tandis que dans le système TDNN, l'EER passe de 21,74% à environ 18%.

Visualisation t-SNE. Comme il a été mentionné, dans le cas du ResNet nous ne pouvions pas trouver un mapping entre les x-vecteurs bruités et les x-vecteurs propres correspondants. Pour tenter de comprendre ce qui se passe, nous avons fait une visualisation des x-vecteurs bruités et propres avec t-SNE (Figure. 1). Nous avons choisi, d'une manière aléatoire, un x-vecteur aléatoire et ses 1000 voisins les plus proches. Les vecteurs choisis sont tracés à côté de leur version propre correspondante. Le t-SNE est formé avec les x-vecteurs propres et les x-vecteurs bruités. Cette visualisation montre, dans le cas du TDNN, deux nuages bien distincts dans l'espace, le premier correspondant aux x-vecteurs propres et le second correspondant aux x-vecteurs bruités. Ce qui nous laisse penser qu'un mapping entre les deux nuages est possible et peut apporter des améliorations significatives. Ce qui est confirmé par les résultats expérimentaux de compensation dans le cas du TDNN. En revanche, dans le cas du ResNet, les deux nuages (x-vecteurs propres et bruités) sont entremêlés dans l'espace. Ce qui suggère une grande robustesse, face aux bruits, des x-vecteurs du ResNet. Par contre, la même constatation nous incite à penser qu'un mapping (avec un DNN ou un modèle statistique) serait plus difficile à trouver et apporterait s'il existe de petites améliorations dans l'absolu. Ce qui est en accord (mais n'explique pas) avec les résultats obtenus avec compensation dans le cadre du ResNet.

TABLE 6 – Résultats après compensation de bruit additif et réverbération (EER)

Système	Propre	Bruité	De-bruité
TDNN	15.21	21.47	18.00
Resnet	6.27	12.48	12.18

Les techniques de débruitage (ou de compensation) tentent de réduire la MSE entre les x-vecteurs bruités et les x-vecteurs propres. Nous avons observé que dans le cas du ResNet, nous avons un petit gain relatif en termes de MSE entre les x-vecteurs bruités et les x-vecteurs débruités, tandis que dans le réseau TDNN, la MSE s'améliore de manière significative. Par exemple, dans le cas du bruit et de la réverbération avec le TDNN, la MSE passe de 0,21 à 0,07 sur l'ensemble de test Fabiol, tandis qu'avec le ResNet la MSE passe de 4,18 à 3,94 en utilisant un DAE. Ce petit gain relatif de MSE ne conduit qu'à une très petite amélioration de l'EER (voir Table 6).



(a) TDNN sans compensation (b) TDNN avec compensation (b) ResNet sans compensation

FIGURE 1 – Visualisation t-SNE du TDNN et du ResNet

6 Conclusion

Dans cet article, nous avons exploré la robustesse face aux bruits de deux systèmes de reconnaissance du locuteur de l'état-de-l'art, le TDNN et le ResNet. Nous avons montré à travers nos expériences que le système basé sur ResNet est nettement plus robuste aux bruits (bruits additif et réverbération) que le le TDNN, même s'il subsiste une dégradation liée aux bruits. Le résultat le plus inattendu est que les techniques de compensation (à base de DAE ou de i-MAP) ne permettent qu'une amélioration marginale dans le cas du ResNet, tandis que l'amélioration est significative dans le cas du TDNN. Malgré ce constat, le système ResNet reste largement plus performant que le TDNN, avec ou sans bruit, avec ou sans compensation. Ces résultats se confirment grâce à une visualisation graphique (t-SNE) des x-vecteurs. L'objectif de futurs travaux est de comprendre le pourquoi de ce comportement du ResNet : atteignons-nous la limite de la compensation ou certaines hypothèses requises par les techniques de compensation ne sont-elles plus vérifiées dans le cas des x-vecteurs extraits par un ResNet ?. Dans ce dernier cas, peut-on imposer la vérification de ces hypothèses lors de l'apprentissage de l'extracteur des x-vecteurs du ResNet ?.

Remerciements

Ce travail a été soutenu financièrement par le projet ANR ROBOVOX.

Références

(2018). D. povey, g. cheng, y. wang, k. li, h. xu, m. yarmohammadi, s. khudanpur, semi-orthogonal low-rank matrix factorization for deep neura lnetworks. p. 3743–3747.

ARSHA NAGRANI, JOON SON CHUNG W. X. A. Z. (2020). Voxceleb : Large-scale speaker verification in the wild. volume Computer Speech Language.

BRIGNATZ V., DURET J. M. D. R. M. (2021). Language adaptation for speaker recognition systems using contrastive learning. In *SPECOM 2021. Lecture Notes in Computer Science*, volume 12997.

COLLEEN RICHEY, MARIA A.BARRIOS Z. A. C. B. H. F. M. G. A. L. M. K. N. A. S. J. V. H. P. G. J. H. C. S. K. N. (2018). Voices obscured in complex environmental settings (voices).

D. CAI Z. C. & LI M. (2018). Deep speaker embeddings with convolutional neural network on supervector for text-independent speaker recognition (apsipa asc).

D. SNYDER, D. GARCIA-ROMERO G. S. A. M. D. P. S. K.-D. (2019). Speaker recognition for multi-speaker conversations using x-vectors. In *IEEE International Conference onAcoustics, Speech and Signal Processing (ICASSP)*.

D. SNYDER, D. GARCIA-ROMERO G. S. D. P. S. K. (2015). Musan : A music, speech, and noise corpus.

HE K., ZHANG X., REN S. & SUN J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 770–778.

K. HE, X. ZHANG S. R. J. S. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision andpattern recognition*, p. 770–778.

LAURENS VAN DER MAATEN G. H. (2008). Visualizing data using t-sne. In *Journal of Machine Learning Research*, p. 2579–2605.

MINH PHAM Z. L. & WHITEHILL J. (2020). How does label noise affect the quality of speaker embeddings ? In *INTERSPEECH*.

MOHAMMADAMINI, MOHAMMAD D. M. J.-F. B. R. S. S. D. D. J. (2021a). Compensate multiple distortions for speaker recognition systems. volume EUSIPCO.

MOHAMMADAMINI, M. M. D. (2021b). Data augmentation versus noise compensation for x-vector speaker recognition systems in noisy environments. volume 2020 28th European Signal Processing Conference (EUSIPCO), 2021.

MOHAMMADAMINI, M. M. D. N.-P. (2020). Denoising x-vectors for robust speaker recognition. In *Proc. Odyssey 2020 The Speaker and Language Recognition Workshop*, p. 75–80.

NOISE B. (2018). Bbc noise. In *BBC*.

SNYDER D., GARCIA-ROMERO D., SELL G., POVEY D. & KHUDANPUR S. (2018). X-vectors : Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 5329–5333.

SUWON SHON, HAO TANG J. G. (2019). Voiceid loss : Speech enhancement for speaker verification. In *INTERSPEECH*.

WAAD BEN KHEDER, DRISS MATROUF. PIERRE-MICHEL BOUSQUET. JEAN-FRANÇOIS BONASTRE M. A. (2017). Fast i-vector denoising using map estimation and a noise distributions database for robust speaker recognition. In *Computer Speech Language*, volume 45, p. 104–122.

ZEINALI, H. W. S. S. A.-M. P. P.-O. (2019). But system descriptionto voxceleb speaker recognition challenge 2019.

ZHONGXIN BAI X.-L. Z. (2021). Speaker recognition based on deep learning : An overview. In *Neural Networks*, volume 140.