Avrupa Bilim ve Teknoloji Dergisi
Özel Sayı 26, S. 217-222, Temmuz 2021
© Telif hakkı EJOSAT'a aittir
Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
Special Issue 26, pp. 217-222, July 2021
Copyright © 2021 EJOSAT
Research Article

A Comparative Assessment of Text-independent Automatic Speaker Identification Methods Using Limited Data

Mandana Fasounaki*, Emirhan Burak Yüce[2], Serkan Öncül[2], Gökhan İnce[1,4]

[1] İstanbul Teknik Üniversitesi, Bilgisayar Mühendisliği, Fakültesi, Bilgisayar Mühendisliği Bölümü, İstanbul, Türkiye (ORCID: 0000-0002-0034-030X)
[2] Arçelik Araştırma ve Geliştirme Merkezi (ORCID: 0000-0002-4220-9940)
[3] Arçelik Araştırma ve Geliştirme Merkezi (ORCID: 0000-0001-9302-8712)
[4]AI and Data Science Application and Research Center

(International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA) 2021 – 11-13 June 2021)

(**DOI:** 10.31590/ejosat.950218)

ATIF/REFERENCE: Fasounaki, M. & Yüce, E. & Öncül S. & İnce G. (2021). A Comparative Assessment of Text-independent Automatic Speaker Identification Methods Using Limited Data. *European Journal of Science and Technology*, (26), 217-222.

Abstract

Automatic Speaker Identification (ASI) is one of the active fields of research in signal processing. Various machine learning algorithms have been used for this purpose. With the recent developments in hardware technologies and data accumulation, Deep Learning (DL) methods have become the new state-of-the-art approach in several classification and identification tasks. In this paper, we evaluate the performance of traditional methods such as Gaussian Mixture Model-Universal Background Model (GMM-UBM) and DL-based techniques such as Factorized Time-Delay Neural Network (FTDNN) and Convolutional Neural Networks (CNN) for text-independent closed-set automatic speaker identification on two datasets with different conditions. LibriSpeech is one of the experimental datasets, which consists of clean audio signals from audiobooks, collected from a large number of speakers. The other dataset was collected and prepared by us, which has rather limited speech data with low signal-to-noise-ratio from real-life conversations of customers with the agents in a call center. The duration of the speech signals in the query phase is an important factor affecting the performances of ASI methods. In this work, a CNN architecture is proposed for automatic speaker identification from short speech segments. The architecture design aims at capturing the temporal nature of speech signal in an optimum convolutional neural network with low number of parameters compared to the well-known CNN architectures. We show that the proposed CNN-based algorithm performs better on the large and clean dataset, whereas on the other dataset with limited amount of data, traditional method outperforms all DL approaches. The achieved top-1 accuracy by the proposed model is 99.5% on 1-second voice instances from LibriSpeech dataset.

Keywords: Speaker Identification, Deep Learning, CNN, Signal Processing, GMM-UBM.

Sınırlı Veri Kullanılarak Metinden Bağımsız Otomatik Konuşmacı Tanıma Yöntemlerinin Karşılaştırmalı Bir Değerlendirmesi

Öz

Otomatik Konuşmacı Tanıma, sinyal işlemedeki aktif araştırma alanlarından biridir. Bu amaçla çeşitli makine öğrenme algoritmaları kullanılmıştır. Donanım teknolojilerindeki ve veri birikimindeki son gelişmelerle birlikte, Derin Öğrenme yöntemleri, çeşitli sınıflandırma ve tanımlama görevlerinde en son teknolojiye sahip yeni yaklaşım haline gelmiştir. Bu makalede, metinden bağımsız, kapalı-küme otomatik konuşmacı tanımlama için Gauss Karışım Modeli-Evrensel Arka Plan Modeli (GMM-UBM) gibi geleneksel yöntemlerin ve Faktörize Zaman Gecikmeli Sinir Ağı ve Evrişimli Sinir Ağları gibi derin öğrenme tabanlı tekniklerin performansını değerlendiriyoruz. Bu karşılaştırmalar, farklı koşullara sahip iki veri kümesinde değerlendirildi. Deneysel veri kümelerinden biri LibriSpeech. Bu veri seti çok sayıda konuşmacıdan oluşan sesli kitaplardan toplanan temiz ses sinyallerinden oluşmaktadır. Ayrıca, müşterilerin bir çağrı merkezindeki temsilcilerle doğal konuşmalarından oluşan bir veri kümesi ise bizim tarafımızdan toplandı ve hazırlandı. Çağrı merkezi veri setindeki ses örnekleri sinyal-gürültü oranı düşük ve oldukça sınırlı sayıda ses örnekleri mevut.

* Corresponding Author: Istanbul Teknik Üniversitesi, Bilgisayar Mühendisliği Fakültesi, Bilgisayar Mühendisliği Bölümü, İstanbul, Türkiye, ORCID: 0000-0002-8332-7054, fasounaki17@itu.edu.tr

Konuşmacı sorgulama aşamasındaki konuşma sinyallerinin süresi, otomatik konuşmacı tanımlama yöntemlerinin performanslarını etkileyen önemli bir faktördür. Bu çalışmada, kısa konuşma bölütlerinden otomatik konuşmacı tanımlaması için bir CNN mimarisi önerilmiştir. Mimari tasarımı, iyi bilinen CNN mimarilerine kıyasla düşük sayıda parametre ile optimum bir evrişimsel sinir ağıdır ve konuşma sinyalinin zamansal yapısını yakalamayı amaçlamaktadır. Önerilen CNN tabanlı algoritmanın büyük ve temiz veri setinde daha iyi performans gösterdiğini, buna karşın sınırlı miktarda veriye sahip diğer veri setinde geleneksel yöntemin tüm derin öğrenme yaklaşımlarından daha iyi performans gösterdiğini gözlemledik. Önerilen model tarafından elde edilen doğruluk, LibriSpeech veri setinden 1 saniyelik ses örneklerinde %99,5'tir.

Anahtar Kelimeler: Konuşmacı Tanımlama, Derin Öğrenme, Evrişimsel Sinir Ağları, Sinyal İşleme GMM-UBM.

1. Introduction

It is known that human voice has some inherent characteristics that make it possible to discriminate them from each other. These features are produced by human vocal system, which has a unique structure in each person. When an individual hears someone speaking and gets familiar with that person's voice, speech flow, etc., he/she can re-identify that person in the future based on the prior knowledge (Latinus and Belin, 2011).

Human voice is one of the biometrics like fingerprint, iris or DNA (Jain et al., 2007). Therefore, it is used for user authentication in various systems and surveillance purposes. During speaking, humans use their voices in different ways, which makes person re-identification based on voice a complicated problem compared to that with other biometrics such as fingerprint and DNA, which are considered static features. Extensive research has been conducted in the field of Automatic Speaker Identification (ASI) since the 1960s (Wolf, 1969; Soong et al., 1987) Generally, ASI is categorized into two types; text-dependent and text-independent. In text-dependent ASI, the spoken utterances are pre-determined. In text-independent speaker identification, there is no constraint on what is being said by the speaker. In our work, we focus on text-independent closed-set speaker identification. In closed-set ASI, it is assumed that the test audio signals belong to the enrolled speakers.

As a pioneering study, Gaussian Mixture Model-Universal Background Model (GMM-UBM) for text-independent speaker identification was introduced by Reynolds (Reynolds, 1992). GMM-UBM has become the most popular ASI method for decades (Zheng et al., 2004; Chowdhury et al., 2010). UBM is a GMM, trained on a large dataset of speech. The parameters of the model are then adapted to speaker-dependent characteristics using the enrollment data with Maximum A Posterior (MAP) parameter estimation. The performance of this method, however, decreases with the increasing the number of speakers and inter-session variability.

Joint Factor Analysis (JFA) aims at solving this problem by factorizing the input signal into speaker-dependent, speaker-independent, channel-dependent and residual factors (Kenny, 2005; Kenny et al., 2014). Also, i-vector approach was introduced for ASI, which is a simplified version of JFA (Kanagasundaram et al., 2011). In i-vector based models, a UBM is trained on a large speech dataset. Using the statistics of the UBM and total variability vector, i-vectors are obtained for each test segment. I-vectors are classified using Linear Discriminant Analysis (LDA) (Jin and Waibel, 2000) or Probabilistic Linear Discriminant Analysis (PLDA) (Kanagasundaram et al., 2012; Senoussaoui et al.,2011).

Recently, deep learning methods are ubiquitously utilized for ASI. Among different DL methods, Convolutional Neural Networks (Lukic, 2016; Nagrani et al., 2017), Factorized Time-delay Neural Networks (FTDNN) (Villalba et al., 2020) and deep metric learning approaches such as Siamese and Triplet Networks (Chung et al., 2020) are the most promising techniques. It is known that DL methods require a large amount of data to work efficiently.

In this paper, we present a Deep Neural Network (DNN) architecture for automatic speaker identification from short utterances. Our model is a Convolutional Neural Network (CNN) with rectangular kernels in the first layer that capture the temporal characteristics of the speech. In order to show the performance of the proposed model, we assess it in several experiments using a small dataset with degraded speech sounds and a large database with clean speech recordings. Furthermore, its performance is benchmarked against the performance of a statistical machine learning technique, GMM-UBM, and other Deep Learning (DL) based methods, such as FTDNN.

2. Automatic Speaker Identification Systems

In this section, main ASI methods that have been examined in this paper are described. One of them is GMM-UBM, which is a traditional ASI approach, and the others are DL techniques, i.e., CNN and FTDNN-based methods.

2.1. Feature Extraction

Most of the speaker identification systems use Mel Frequency Cepstral Coefficients (MFCC) as input acoustic features. Mel scale describes the perceptual distance between pitches of different frequencies, and is known to imitate the logarithmic perception of human auditory system (Beigi, 2011). In order to extract MFCC from speech data, pre-emphasis is applied to the signal. Then the signal is framed, and Hanning window is applied to each frame. Fast Fourier Transform (FFT) is applied to the windowed frames. By calculating the log of power spectral density, and applying Discrete Cosine Transform (DCT) to the magnitude, MFCCs are extracted.

2.1.1. Gaussian Mixture Model-based ASI

Gaussian Mixture Model (GMM) is a data clustering algorithm, mostly used for finding the distribution of subpopulations in a population. The assumption is that the data is normally distributed. GMMs are trained using Expectation- Maximization algorithm (Moon, 1996), which is an iterative algorithm to produce maximum a posteriori estimates of a statistical model.

In closed-set speaker identification, during the training phase, a UBM is trained using a large voice dataset consisting of different data to model the speaker-independent characteristics of acoustic features. The UBM's parameters are adapted to the enrollment data by MAP estimation. During the test phase, log-likelihoods of the test segment for all GMMs are computed. The scores are normalized, and the model with the highest likelihood score is selected as the target speaker.

2.1.2. Deep Learning-based ASI

1) Convolutional Neural Network-based ASI: CNN-based methods were initially designed for image classification (Krizhevsky et al., 2012). In these models, the convolution filters are square kernels with sizes (3×3), (5×5), (7×7), etc. These architectures support the spatial characteristics of the input and mostly ignore temporal information. Considering the time-related nature of speech, we modified the conventional CNN architecture, so that it supports the temporal information of speech signals.

The first layer of the network contains several rectangular filters that cover all the features of consecutive frames. The size of the filters are (#of features × m), where m is the number of frames that are supported by the kernels. The following convolutional layers have filters of sizes (1 × x), where x differs based on the output size of the previous layers. As shown in Figure 1, the output is narrowed in each layer. Before the classification part of the model, the output of the last convolutional layer is flattened into a fixed

size embedding. The subsequent fully connected (FC) layers classify the embedding into one of the speakers categorizes. The identity functions or skip connections in the classifier help preserving the information from previous FC layers

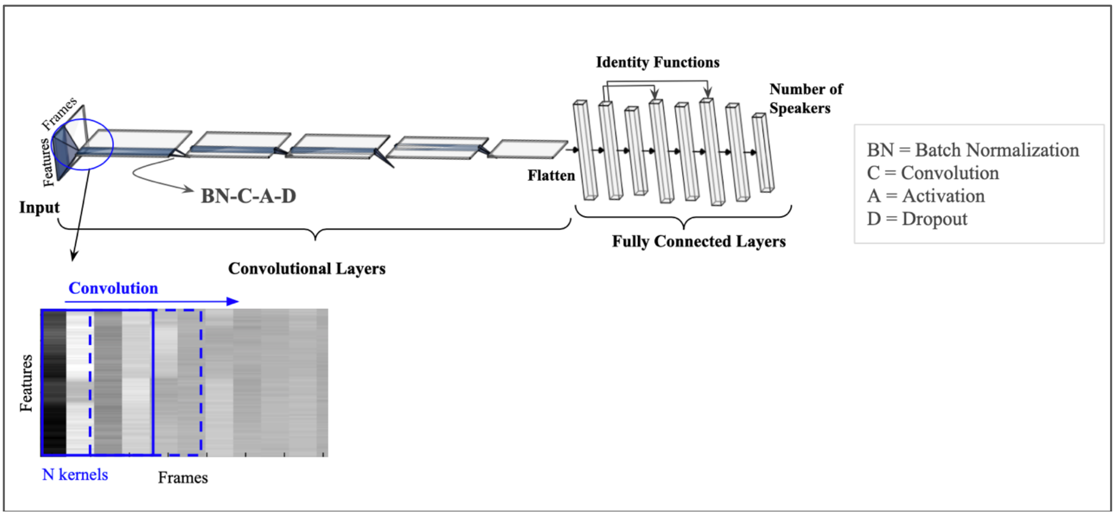

Figure 1: The Proposed CNN Architecture

2) Factorized Time-delay Neural Network-based ASI: The model setup is the same as in (Villalba et al., 2020). 1-D convolutional kernels are applied on frame level, in each layer of the network. By calculating mean and variance of the output of the final layer and aggregating them, embeddings are obtained. These embeddings are called x-vector. The x-vectors are classified by a two-layered neural network.

3. Experiments

3.1. Datasets

In this work, two main datasets were used to evaluate the performance of different methods. Librispeech (Panayotov et al., 2015) is a clean dataset with large amount of data, unlike the speech data gathered from a call center specifically for this study. The details of the datasets are shown in Table 1. In both datasets the number of male and female speakers is balanced.

Table 1. Statistics of the Experimental Datasets

Dataset	**# of Speakers**	**Per-speaker Length**	**Total Hours**	**# of Instances per speaker**
Call Center	411	29 seconds	4	11 ± 5
LibriSpeech	251	25 minutes	100	103 ± 15

3.1.1. Call Center Dataset

This dataset was created using a call center's telephone conversations of Turkish speakers. The speech data consists of natural conversations of customers with call center agents, therefore they are subject to effects caused by different mobile phones, microphones, environmental settings, and multiple simultaneous speakers. Also, sampling rate is 8kHz because the call centers tend to save the conversations in a way that reduces the stored data size.

To prepare this dataset, 750 speakers' conversations were inspected. The speakers with less than 25 seconds of speech were eliminated. Also, the recordings with more than one speaker, extreme external noise, or deteriorated voice signals were removed. In the final dataset, there are 411 speakers, where each speaker class contains the average of 11 utterances taking 29 seconds. For cleaning out the dataset, we excluded the silent parts of the conversations. Overall, the dataset comprises four hours of speech data that can be considered "limited data" for an ASI task.

Because of the limited data at hand, we isolated 3 speech instances with the average duration of 3 seconds for each speaker to constitute our test set. The remaining portion of the dataset is used as enrollment set. 80% of the enrollment data is used for training and 20% for evaluating the model.

3.1.2. LibriSpeech Dataset

LibriSpeech is a large dataset consisting of read English speech derived from audiobooks. This dataset contains more than 1000 hours of speech, sampled at 16 kHz. In our experiments, we use a subset of LibriSpeech, which contains 100 hours of speech for 251 speakers, with 103 utterances taking around 25 minutes in total for each speaker in average.

For creating the test set, we separated 10 speech instances with the average length of 15 seconds for each speaker. The remaining portion of the dataset is used for training.

3.2. Experimental Setup

In the preprocessing step, the speech signal is framed into 25-millisecond frames, with 15 ms overlap. We used 13 MFCCs extracted from each frame. 1-second speech segments create (13 ×100) dimensional inputs.

3.2.1. GMM-UBM Settings

In our experiments, a UBM was trained on a large number of speech voices, for 100 iterations. The number of components in UBM was 1024 and the covariance matrices are diagonal matrices. The parameters of UBM are adapted to the enrollment dataset using MAP estimation, to create individual GMM for each speaker. In the test phase, for each test segment, log-likelihoods of all the GMMs are computed. After the score normalization, the speaker model with the highest likelihood score is accepted as the target speaker.

3.2.1. Proposed CNN Settings

In Table 2, the sizes of convolutional filters in the proposed CNN are shown for 1-secondmspeech signals with 13 MFCCs.

Table 2. Kernel Sizes in the Proposed CNN

Layer	**Kernel Size**	**# of Kernels**
Input	(13, 100)	-
1	(13, 20)	1024
2	(1, 20)	512
3	(1, 16)	512
4	(1, 16)	512
5	(1, 10)	128

Table 3 contains the details of the fully connected layers in the classifier part of the network. Sigmoid function is used after all layers except for the last layer. The activation function of the output layer is Softmax.

Table 3. Fully Connected Part of the Proposed CNN

Layer	**Layer Size**	**Skip Connection**
6	1024	-
7	512	-
8	512	6
9	128	-
10	128	6
11	512	-
12	# of Speakers	-

In the proposed CNN, batch normalization was applied before each convolution layer, including the first layer. High dropout rates (0.5, 0.6) were used in all layers, which significantly improved the performance. With heuristic search, we obtained the optimal hyperparameters for the model. Loss function was selected as categorical Cross Entropy, and the optimizer was determined to be Adam with learning rate of 0.001. The optimal batch size was 128. The model was trained using backpropagation for 40 epochs.

The number of parameters in our model is 4.2 million, which shows the network is more efficient compared to the networks that are used for ASI, i.e., VGGNet (Simonyan K., 2014) having approximately 138 million parameters.

We used Python and Tensorflow for implementation, and trained the model on AWS Deep Learning AMI. The metric for evaluating the methods is accuracy as in Eq. (1).

$$Accuracy = \frac{\#\, of\ correct\ detections}{\#\, of\ all\ test\ instances} \times 100 \qquad (1)$$

Top-1 accuracy shows the percentage of test samples that were correctly identified by the model. Top-5 accuracy indicates the percentage of the data samples, where the target speaker is in the first five predictions.

4. Results

In Table 4, the achieved accuracies by different methods on LibriSpeech and Call Center dataset are shown. GMM-UBM and FTDNN are used for benchmarking purposes.

Table 4. Performances of Different Methods on 1-second Voice Segments

Dataset	**Method**	**Top-1 Accuracy [%]**	**Top-5 Accuracy [%]**
LibriSpeech	GMM-UBM	97.3	98.3
	Proposed CNN	**99.5**	**99.8**
	FTDNN	96.6	97.5
Call Center	GMM-UBM	**62.5**	**80.4**
	Proposed CNN	30.2	60.1
	FTDNN	20.7	30.4

All competing methods, including FTDNN and GMM-UBM show high accuracy using utterances with the duration of 1 second, while CNN outperformed both the GMM and FTDNN when clean and long speech recordings as in LibriSpeech are used. It is shown that the proposed architecture, CNN-based ASI, achieves 99.5% accuracy on 1-second segments of utterances.

However, the results drop significantly when the Call Center Dataset is used. Both the DL-based ASI approaches fail in successfully identifying the speakers when noisy and low-quality sounds were utilized for training. The accuracy of the proposed CNN approach is as low as 30.2%, whereas the GMM demonstrates a drastically better performance than CNN (62.5%). On both datasets, FTDNN shows inferior performance.

Training duration in DL-based methods are generally higher than the traditional approaches. FTDNN was trained for 54 hours on LibriSpeech. With the proposed CNN, the training was completed in 2.5 hours. Both models were trained for 40 epochs on NVIDIA Tesla K80. Training time for GMM is 30 minutes on Intel Xeon E5-2686 v4 @2.30GHz.

In order to investigate the performance of GMM more elaborately, in Table 5, the accuracy of GMM-based ASI with respect to the number of speakers in the dataset is observed.

Table 5. Accuracy of GMM-UBM on Call Center Dataset

Number of Speakers	**Top-1 Accuracy**
40	89.5
60	84.4
180	82.5
250	66.8
411	**62.5**
500	62.3

It is observed that in the limited and noisy dataset, GMMUBM perform better with smaller number of speakers. By increasing the number of speakers in Call Center dataset, the performance decreases.

5. Conclusion

In this paper, we presented a modified version of conventional convolutional neural networks for text-independent speaker identification using short utterances. We assessed the performance of different ASI methods on two datasets; one dataset with sparse and noisy data gathered from telephone conversations with a call center, and a large speech corpus called LibriSpeech. With abundant data, the proposed CNN outperforms the state-of-the-art methods (e.g., FTDNN) and traditional methods (e.g., GMM-UBM). Most of the existing systems use 3-second or longer utterances for depicting an acceptable performance (Li et al., 2020). However, we showed that our model achieves 99.5% accuracy even on 1-second speech segments.

We conclude that in the applications that support limited number of users and require fast response/decision making, GMM-UBM performs better compared to DL methods. Also, for fast enrollment of new speakers to an ASI system, GMM-UBM is definitely easier to manage and computationally cheaper, thus more practical. In DL-based method, for adding a new speaker, the model needs to be trained on all data, but in GMM only the new speaker's model is created and inserted to the model base.

For further research, we aim at using noise reduction, speaker diarisation and separation techniques on call center data to improve the performance of GMM-based ASI. Also, channel compensation techniques will be employed to reduce inter-session variability.

6. Acknowledgment

This study has been supported by Arcelik ITU R&D Center and the Scientific Project Unit (BAP) of Istanbul Technical University, project number: MOA-2019-42321. The authors thank to Cagri Aslanbas, Berna Erden, Pinar Baki, Ugur Halatoglu and Baris Bayram for their fruitful discussions.

References

Latinus, M. and Belin, P. (2011). Human voice perception. Current Biology, 21(4):R143 – R145.

Jain, A. K., Flynn, P., and Ross, A. A. (2007). Handbook of biometrics. Springer Science & Business Media.

Beigi, H. (2011). Fundamentals of Speaker Recognition. Springer Publishing Company, Incorporated.

Wolf, J. J. (1969). Acoustic measurements for speaker recognition. The Journal of the Acoustical Society of America, 46(1A):89–90.

Soong, F. K., Rosenberg, A. E., Juang, B., and Rabiner, L. R. (1987). Report: A vector quantization approach to speaker recognition. AT T Technical Journal, 66(2):14–26.

Reynolds, D. (1992). A Gaussian Mixture Modeling Approach to Text-independent Speaker Identification. College of Engineering, Georgia Institute of Technology.

Zheng, R., Zhang, S., and Xu, B. (2004). Text-independent speaker identification using gmm-ubm and frame level likelihood normalization. In 2004 International Symposium on Chinese Spoken Language Processing, pages 289–292. IEEE.

Chowdhury, M. F. R., Selouani, S.-A., and O'Shaughnessy, D. (2010). Text-independent distributed speaker identification and verification using gmm-ubm speaker models for mobile communications. In 10th International Conference on Information Science, Signal Processing and their Applications (ISSPA 2010), pages 57–60. IEEE.

Kenny, P. (2005) Joint factor analysis of speaker and session variability: Theory and algorithms. RIM, Montreal,(Report) CRIM-06/08-13, 14,28-29.

Kenny, P., Stafylakis, T., Ouellet, P., and Alam, M. J. (2014). Jfa-based front ends for speaker recognition. In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1705–1709. IEEE.

Kanagasundaram, A., Vogt, R., Dean, D. B., Sridharan, S., and Mason, M. W. (2011). I-vector based speaker recognition on short utterances. In "Proceedings of the 12th Annual Conference of the International Speech Communication Association", pages 2341–2344. International Speech Communication Association (ISCA).

Jin, Q. and Waibel, A. (2000). Application of lda to speaker recognition. In Sixth International Conference on Spoken Language Processing.

Kanagasundaram, A., Vogt, R. J., Dean, D. B., and Sridharan, S. (2012). Plda based speaker recognition on short utterances. In "The Speaker and Language Recognition Workshop (Odyssey 2012)". ISCA.

Senoussaoui, M., Kenny, P., Brümmer, N., Villiers, E. d., and Dumouchel, P. (2011). Mixture of plda models in ivector space for gender independent speaker recognition. In Twelfth Annual Conference of the International Speech Communication Association.

Y. Lukic, C. Vogt, O. Dürr and T. Stadelmann, "Speaker identification and clustering using convolutional neural networks," *016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP), 2016, pp. 1-6, doi: 10.1109/MLSP.2016.7738816*

Nagrani, A., Chung, J. S., & Zisserman, A. (2017). Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*.

Chung, J. S., Huh, J., Mun, S., Lee, M., Heo, H. S., Choe, S., Ham, C., Jung, S., Lee, B.-J., and Han, I. (2020). In defence of metric learning for speaker recognition. arXiv preprint arXiv:2003.11982.

Villalba, J., Chen, N., Snyder, D., Garcia-Romero, D., McCree, A., Sell, G., Borgstrom, J., García-Perera, L. P., Richardson, F., Dehak, R., Torres-Carrasquillo, P. A., and Dehak, N. (2020). State-of-the-art speaker recognition with neural network embeddings in nist sre18 and speakers in the wild evaluations. Computer Speech & Language, 60:101026.

Moon, T. K. (1996). The expectation-maximization algorithm. IEEE Signal Processing Magazine, 13(6):47–60.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, NIPS’12, pages 1097–1105, USA. Curran Associates Inc.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech: an asr corpus based on public domain audio books. In 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP), pages 5206–5210. IEEE.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Li, R., Jiang, J.-Y., Li, J. L., Hsieh, C.-C., and Wang, W. (2020). Automatic speaker recognition with limited data. In Proceedings of the 13th International Conference on Web Search and Data Mining, pages 340–348.