Investigating the use of multiple languages for crisp and fuzzy speaker identification

DA COSTA ABREU, Marjory <http://orcid.org/0000-0001-7461-7570> and AGUIAR DE LIMA, T.

Investigating the use of multiple languages for crisp and fuzzy speaker identification

Thales Aguiar de Lima[1] **and Márjory Da Costa-Abreu**[2]

[1]Center of Earth and Exact Sciences, Federal University of Rio Grande do Norte, Natal, BR
[2] Department of Computing, Sheffield Hallam University, Sheffield, U.K.

Keywords: speaker identification, multilingual, Brazilian Portuguese, fuzzy.

Abstract

The use of speech for system identification is an important and relevant topic. There are several ways of doing it, but most are dependent on the language the user speaks. However, if the idea is to create an all-inclusive and reliable system that uses speech as its input, we must take into account that people can and will speak different languages and have different accents. Thus, this research evaluates speaker identification systems on a multilingual setup. Our experiments are performed using three widely spoken languages which are Portuguese, English, and Chinese. Initial tests indicated the systems have certain robustness on multiple languages. Results with more languages decreases our accuracy, but our investigation suggests these impacts are related to the number of classes.

1 Introduction

Speech exists with the main reason to enable communication between humans. This communication translates into a sequence-dependent and rule-based system that we call *language*. To talk with one another, humans use a complex system to produce the voice signal. Starting at the lungs, through the trachea, stimulating vocal cords and the larynx tube, using the pharynx cavity, the tongue, vellum, mouth and nasal cavity to produce sound finally. This procedure is detailed in [1, p. 5].

Speaker identification (SPiD) is a biometric branch of Automatic Speech Recognition field. This sub-field of speech research focuses on *identity recognition*. A better definition would be "deciding if a speaker is a specific person or is among a group of persons." [2]. On the other hand, speaker verification is "[. . .] deciding if a speaker is whom he claims to be." [2]. This problem can be further specified as *open-set* when the unknown speaker is not enrolled in the system, and as *closed-set* when everyone is registered [3]. Then, some systems rely on the content of the signal, that is, a type of passphrase. Those are classified as *text-dependent*, in contrast to *text-independent* when the user can speak anything [4]. In this paper, we explore **closed-set text-independent speaker identification** systems.

The literature has a great variety for this biometry. When representing a speaker, the Mel-Frequency Cepstrum Coefficients (MFCC) and some variations are still widely adopted [5, 6, 7, 8, 9, 10, 11], even though the state-of-the-art has shifted from it to i-vectors [10, 12] and then towards x-vectors [13].

Besides biometric features, classification has also improved for SPiD. For long, GMM-UBM [4] and HMM [14] dominated the field. However, other methods such as vector quantisation (clustering) [15, 16] have their spots. Little research is made for fuzzy classification [17, 18, 8, 19, 20], but the majority is quite dubious when describing their methods for both models and data. Furthermore, most recent research has converged to Neural Network variants, such as Deep Neural Networks [21, 12, 5], Convolutional Neural Networks [22], and others [23, 24]. Meanwhile, the SPiD community has always speculated the impact of language for the problem [25]. In fact, some studies investigate this topic [26, 27] but they usually employ languages with common ancestry or accent variations.

However, most of these works have a small dataset or do not provide a better description of how to split the dataset or any statistical tests performed. Also, not much research has been done for fuzzy models, even though they have provided decent results. Moreover, only [28] has considered Brazilian Portuguese (BP) in their open-set classification. The low occurrence of this language is due to its lack of resource for creating speech technologies, as explored in our previous work [29]. Therefore, we propose a method to verify how different languages (on structure, accent, and ancestry), BP, English (EN), and Chinese (CN), can affect fuzzy and typical classifier for the SPiD.

Our experiments are conducted on three databases, using different classifiers. Then, after selecting the model with the best performance for BP we add other languages and verify its effects on classification. Results suggest that the typical classification does not suffer from language variation, while decent classification results are obtained for BP. Next, we introduce our methodology describing how we prepare the data for our experiments and our feature extraction setup.

2 Materials and Methods

In this work, we explore three distinct datasets on three different languages. The DARPA-TIMIT [30], the LapsBenchmark16k [31], and the AISHELL-1 [32]. They provide data on EN, BP, and CN, respectively. They were chosen because their audio have equal sampling rate (16KHz), they are public and free. Other available multilingual corpus: NIST SRE datasets,

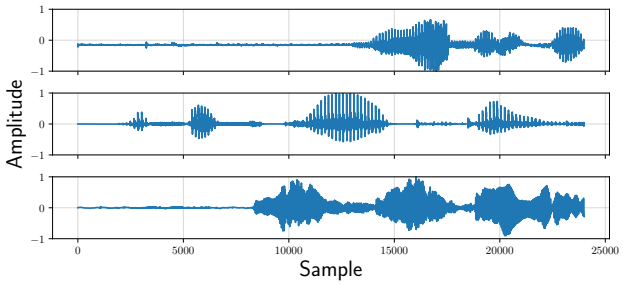


Figure 1: Samples from Portuguese, English, and Chinese (top to bottom).

Call My Net Corpus [33], and more. However, most of them are not free, which puts them over the budget of this research. Each dataset is better detailed in their respective references. Therefore, we limit ourselves to briefly describe them.

The DARPA-TIMIT is a free version of TIMIT. Recordings in this dataset have 2.9s±0.8s of duration. The BP data from LBM16K has 20 recordings per speaker, while their durations are about 4.6s±0.8s. Finally, the AISHELL-1 provides a substantial amount of CN speech. With at least 300 samples per speaker, and an approximated duration of 4.6s±1.3s. Figure 1 shows examples of different samples from each dataset.

Following, Table 1 summarises the main characteristics of the datasets. The gender distribution from BP and EN are not good compared to CN. However, since our goal is to investigate multilingual SPiD, then overall gender is well distributed. Also, the number of samples for each language is different. However, they are balanced by under-sampling.

Dataset	#Size	#Speakers	Gender (M/F)	Lang
DARPA	6,300	630	70%/30%	EN
LBM16K	700	35	72%/28%	BP
AISHELL-1	141,200	400	47%/53%	CN
Total	148,200	1,065	48%/52%	—

Table 1: Summary of the datasets.

This section introduced the data adopted for this work and a portion of the preparation for executing the experiments. Following, we introduce the classification methods.

3 Experimental Setup

This section presents the organisation of our data to test our hypothesis, feature methodologies, and their respective settings. Subsequent procedures were executed on a system with an AMD Ryzen-5 1600 Six-Core Processor, Dual Channel 2×8GiB DIMM DDR4 2400MHz, SSD Kingston A1000 R1500Mb/s and W500Mb/s, [MSI] Radeon RX580 8G OC, 64-bit Pop!_OS 20.04 with Gnome 3.36.2.

First, we test the models with BP, then the EN speakers are added, followed by CN. However, to reduce the number of tests and for better visualisation, the BP+EN and BP+EN+CN are tested only with the best BP classifier. Also, it is crucial to pay attention to the growing number of classes as new languages are added. To assess this problem, we perform 30 experiments with around 34 classes from mixed languages. For that, we randomly select $^1/_3$ of data from each language (Figure 2b), this time ignoring other characteristics from data.

3.1 Preparing the Data

First, we made sure all languages had similar sizes. Mainly, we preserved gender, number of speakers and accents from datasets. Also, DARPA-TIMIT had the least samples per speaker. Thus, we execute an undersampling on AISHELL-1 and LBM16K making them have 10 samples for each speaker through a Roulette algorithm. Furthermore, BP has 35 classes, much less than other languages. Therefore, some English speaker had to be cut-off from experiments, resulting into 34 speakers. In contrast, for CN which uses its development set with 40 speakers. Figure 2 gives an overview of the distributions.

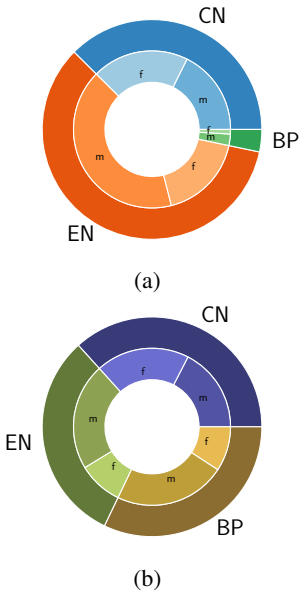
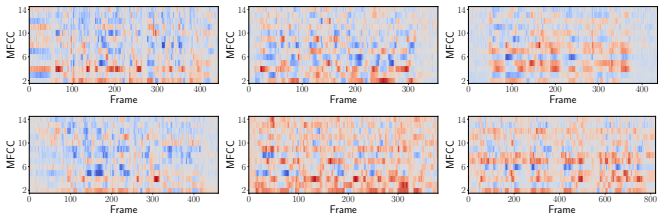


Figure 2: Original (a) and experimental (b) distribution.

3.2 Biometric Feature Extraction

Before extraction, the signals pass through an energy-based voice activity detection. Then, we use the first 13 coefficients from 40 MFCC, excluding the 0*th*. Frames are 25ms long with 10ms stride, a Hamming Window function, as well as a 512 point DFFT. Besides, we used 40 triangular filters spamming from 300Hz to 3400Hz. Then, computing the Δ and $\Delta\Delta$, and appending the logarithm energy, creating a 39-dimensional MFCC feature vector. Finally, a cepstral mean subtraction is applied to the data before training/testing to remove channel and recording variations [4, p. 95]. The feature for a single recording can be represented in the spectral form, as shown in Figure 3.



(a) Brazilian (b) English (c) Chinese

Figure 3: MFCC spectrum for speakers on BP, EN, and CN.

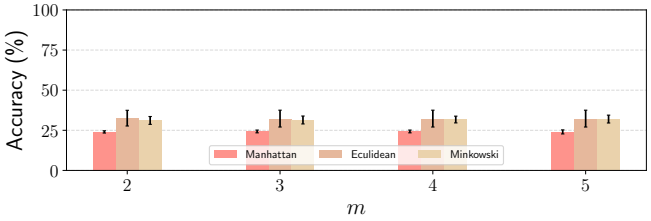


Figure 4: FCM results for speaker recognition on BP.

3.3 Fine Tuning SPiD Systems

Below, we describe parameter search space for our experiments. This procedure was performed with a grid-search implementation available at GitHub[1], along with those hyperparameters and a stratified 3-fold. We choose a stratified version to maintain the class distribution, while a 3-fold guarantee a decent amount of training samples. Furthermore, we use a Fuzzy C-Means (FCM) and Fuzzy K-Nearest Neighbours (FKNN) from GitHub[2], while K-Nearest Neighbours (KNN) and Support Vector Machines (SVM) are of the SKLearn library [34].

FCM have m varying in $\{1.5,\ 2,\ 2.5,\ 3\}$. The number of clusters is fixed at 35, and tolerance at 0.2. Finally, we use the Manhattan, Euclidean and Minkowski similarities.

FKNN have $K \in \{2, 4, \ldots, 12\}$, similarities are Manhattan; Euclidean; and Minkowski, $m \in \{1.5,\ 2,\ 2.5,\ 3\}$. Initialisation search L is fixed to 16.

KNN have variable $K \in \{2, 4, \ldots, 12\}$, distances are Manhattan; Euclidean; Minkowski; and DTW.

SVM has C and γ varying in $\{0.001,\ 0.01,\ 0.1,\ 1,\ 10\}$. The linear, RBF, and sigmoid kernels are tested. We vary the degrees in $\{1, 2, 3, 4, 5\}$ for the polynomial kernel. Residues are not considered.

4 Results

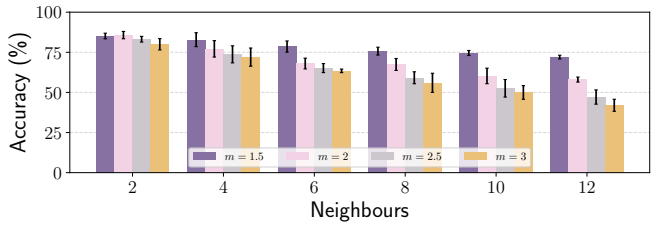
The experimental results are presented here. Not only the precision, but also the performance with respect to languages.

4.1 Fuzzy C-Means

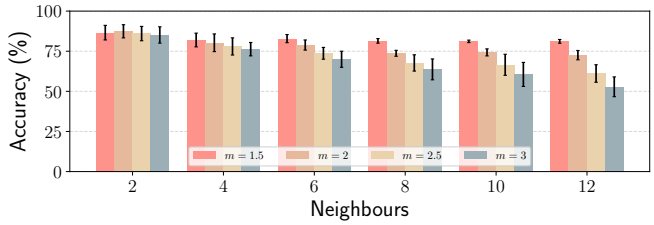
This clustering method reaches a maximum accuracy score of $32.57\% \pm 4.88\%$, illustrated in Figure 4. Setting the fuzziness $m = 1.5$ and the metric to Euclidean produces the best results. Tuning m provided no significant improvement, in contrast to variations on the distances.

4.2 Fuzzy k-Nearest Neighbours

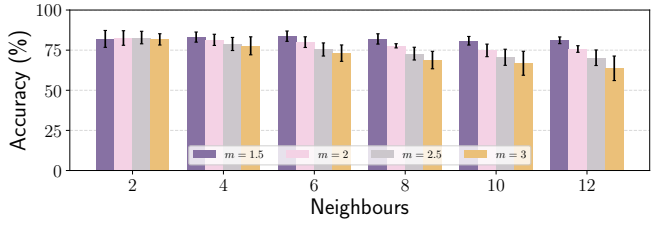
This model had a much better performance, achieving $87.42\% \pm 4.1\%$ accuracy. This score is obtained with Euclidean similarity, 2 neighbours, and $m = 2$. Therefore, it represents an absolute 54.85% improvement over FCM. This and the other scores are presented in Figure 5. Besides, the classification degrades for $m > 2$. This is evident by looking at bars with the same colour at Figure 5. Also, notice that the best setup uses a small k.



(a) Manhattan



(b) Euclidean



(c) Minkowski

Figure 5: Results for Fuzzy k-Nearest Neighbours with different configurations.

4.3 k-Nearest Neighbours

Again, the accuracy is inversely proportional to the number of neighbours (Figure 6).

From Figure 6, the best value is $86\% \pm 3.14\%$, which is achieved by Euclidean, Minkowski, and Manhattan metrics. We take the last metric as the best, as it has the same performance with better generalisation. Therefore, the best setup for KNN is the Euclidean metric with $k = 6$. Also, the KNN results

[1]See: https://github.com/thalesaguiar21/Gryds
[2]See: https://github.com/thalesaguiar21/Fuzzy

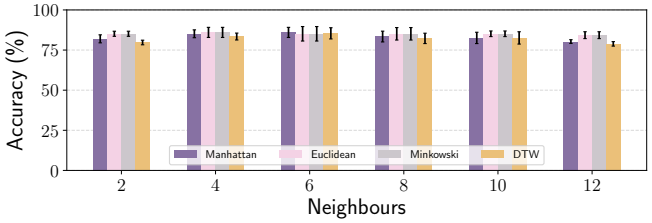
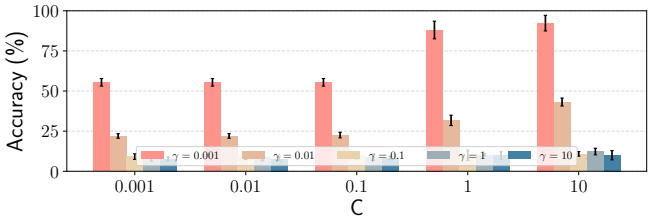


Figure 6: k-Nearest Neighbours accuracy for on BP.

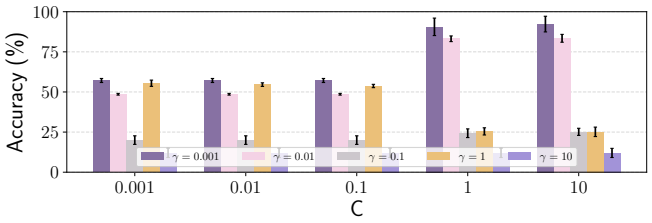
represent an 1.42% attenuation over the FKNN.

4.4 Support Vector Machines

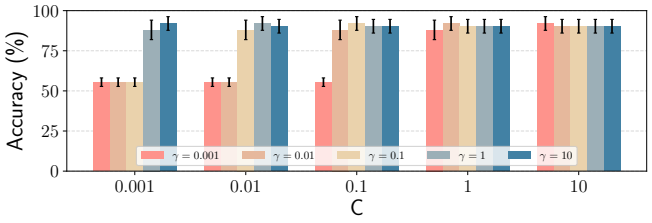
Here, almost every kernel achieves decent accuracy values. Both Sigmoid and RBF get the highest score of 92.29% $\pm$ 4.8%. They reach this value using $\gamma = 10^{-3}$ and $C = 10$. While the Linear kernel is right behind with 92% $\pm$ 4.2% accuracy.



(a) Sigmoid.



(b) RBF.



(c) Linear.

Figure 7: SVM accuracy scores for BP.

Given this tiny difference between them, we decided to further compare these kernels. In short, the Linear kernel has a 7.7% better performance per time, lower σ (Figure 7), smaller test duration, and lower C; while losing in absolute accuracy and γ. Since using a $C = 10$ can lead to a non-generic model, we choose the Linear SVM configuration (SVML1-C01G01). The SVML1-C01G01 is an 4.58% improvement over FKNN. Results for Polynomial kernel are not displayed as their behaviour is very similar to Linear; thus not adding much for the discussion.

4.5 Multilingual Experiments

This section explores the best configuration from monolingual experiments into multilingual environments. Now, SVML1-C01G01 is submitted to experiments with BP+EN and BP+EN+CN. Results are presented in Figures 8 and 9.

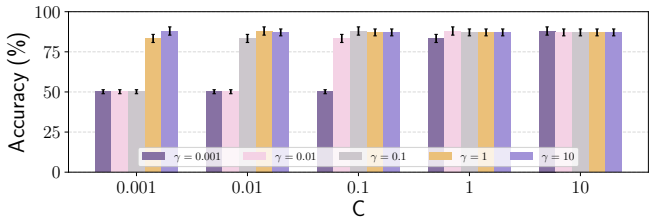


Figure 8: Results of Linear SVM on BP+EN dataset.

Results shown on Figure 8 have a similar behaviour from monolingual SVM tests regarding C with no effect on accuracy for $\gamma \leq 0.01$. The best configuration on this dataset is still $C = 0.01$ and $\gamma = 10^{-1}$, achieving 87.97% $\pm$ 2.56% of accuracy. Therefore, a 4.03% decrease compared to the monolingual experiments. Also, the number of mistakes of BP speakers by EN is small when compared to the opposite, as presented in Figure 10a. Since we carefully extracted em processed our features to remove any bias from languages, recording procedures, or errors added while transforming the signal, these mistakes are more likely related to language distinction than something else.

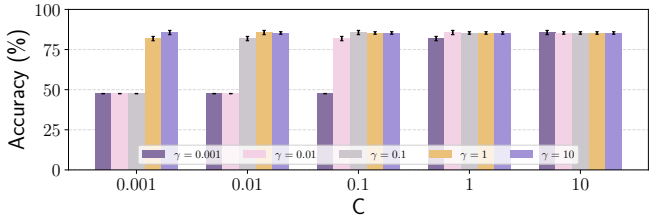
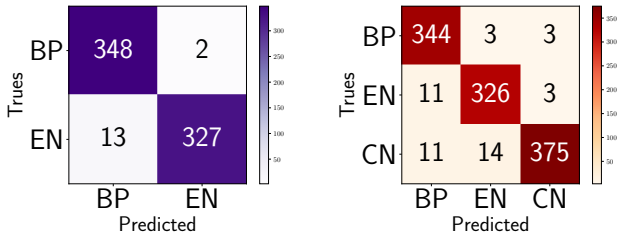


Figure 9: Results of Linear SVM on BP+EN+CN dataset.

Next, adding CN as a third language to the dataset results into 85.59% $\pm$ 1.32% accuracy, a 2.38% decrease compared to BP+EN. Mostly, due to confusions between CN and EN speakers, as shown in Figure 10b. From Figure 10 it is noticeable that there is a few confusions between languages. Besides that, from a total of 45 wrong classifications, 31% (28) are BP, 34% (31) EN, and 45% CN. Next, a total of 30 tests are evaluated using the same configuration from previous multilingual experiments with smaller versions of the multilingual dataset. As average, these experiments achieved 91.88% $\pm$ 1.87%, 0.12% lower than monolingual results.



(a) (b)

Figure 10: Language × Language confusion matrix.

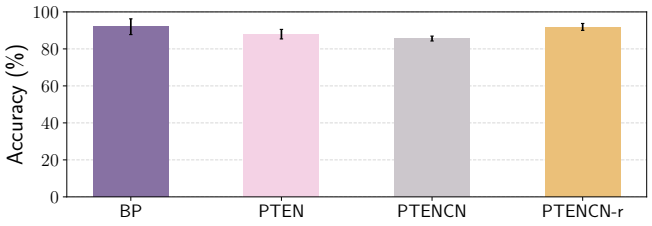


Figure 11: Results for randomly selecting $^1/_3$ classes of each language.

5 Discussion

Here, we discuss and compare the monolingual and multilingual results. Using this discussion to verify our hypothesis that **closed-set speaker identification is language independent**. Our focus is to describe the results with respect to languages, rather than models performance. Even tough, we add a short discussion of our fuzzy results.

The first model we evaluated was the FCM. It was expected to have a bad performance, but the results were surprising. Then, FKNN provides a substantial improvement over the FCM. While followed by a small attenuation from its crisp version, the KNN. However, SVM provided the best performance.

Then, the FKNN had a decent score, surpassing its crisp version by 1.42%. Even though not by a large margin, this value can still improve with further tuning. Here, several membership functions could be tested to better balance the model according to the data. Thus, showing that fuzzy models can be as accurate as their crisp versions; as well as their flexibility.

Results showed that adding a second language reduced the model accuracy by 4.03%, and 6.41% when adding a third one. Thus, indicating that our hypothesis would fail. However, language is not the only variable to consider due to other characteristics that can influence SPiD results. Our results can easily be influenced by gender and number of speakers. From a total of 45 wrong classifications, 31% (28) are BP, 34% (31) EN, and 45% CN; thus representing no bias. The same way for gender, as male speakers appearing on 46% of them. Thus, no problems are found when looking at genders for our results. Therefore, discarding any influence from it.

However, some information from it is very interesting. CN speakers hold 79.16% of opposite gender mistakes, that is, predicting a male speaker with a female or the other way around. From this proportion, the female speakers represent a large amount. Except by 1 test, every mistake of Chinese speaker by English speaker is between female (CN) and male (EN). For Portuguese, from a total of 9 Chinese females, only two are predicted as a female BP speaker. This suggests that Chinese males voices are very distinct from both BP and EN males. While CN females have a close relation to male voices.

Finally, we conducted experiments to assess the increasing number of classes. By discarding gender distribution, we created smaller datasets with approximately 34 speakers, 1 less than monolingual size. Then, these experiments resulted into 91.88% accuracy. Figure 11 compares our main results from each dataset showing that result with our monolingual are only 0.12% above. A fairly close score; thus indicating that consecutive reductions of accuracy in our results are likely due to the increase in classes.

6 Conclusions

This paper presented results for **closed-set text-independent SPiD** for multiple languages. It is crucial to keep in mind that our objective was not achieving high accuracy. In this work, we aim to investigate how SPiD systems performs in multilingual environments. Our results, using the settings employed in this work, languages have little influence on the system accuracy.

Some segments of this work can be improved or expanded. First, most of our findings come to the conclusion that the speaker identification system is language independent, but the influence of the features are not investigated. A comparison between different features, such as x-vectors or LPC could enrich the discussion around multilingual SPiD. Furthermore, a better method to evaluate the influence of the number of classes could be used. These results were obtained through random experiments. A better method would be to split and label each language data, then test all its combinations. This way, one can ensure that all speakers evaluated.

Acknowledgements

This work was supported by the Brazilian Coordination for the Improvement of Higher Education Personnel (CAPES).

References

[1] L. R. Rabiner and R. W. Schafer, *Theory and applications of digital speech processing*, vol. 64. Pearson Upper Saddle River, NJ, 2011.

[2] J. P. Campbell, "Speaker recognition: a tutorial," *Proc IEEE Inst Electr Electron Eng*, vol. 85, pp. 1437–1462, Sept. 1997.

[3] R. Hu and R. I. Damper, "Fusion of two classifiers for speaker identification: removing and not removing silence," in *Proc. 7th Int. Conf. Inf. Fusion*, vol. 1 of *FUSION*, (Philadelphia, PA, USA), pp. 429–436, IEEE, July 2005.

[4] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Commun*, vol. 17, pp. 91–108, Aug. 1995.

[5] P. Anand, A. K. Singh, S. Srivastava, and B. Lall, "Few Shot Speaker Recognition using Deep Neural Networks," Apr. 2019.

[6] A. Basu, S. Bose, A. Pal, A. Mukherjee, and D. Das, "A Novel Minimum Divergence Approach to Robust Speaker Identification." arXiv:1512.05073, Dec. 2015.

[7] K. W. Cheuk, B. T. Balamurali, G. Roig, and D. Herremans, "Latent Space Representation for Multi-Target Speaker Detection and Identification with a Sparse Dataset Using Triplet Neural Networks," in *Proc. ASRU*, ASRU, (SG, Singapore), pp. 358–364, IEEE, Dec. 2019.

[8] A. K. Devika, M. G. Sumithra, and A. K. Deepika, "A fuzzy-GMM classifier for multilingual speaker identification," in *Proc. Int. Conf. Commu. Signal Process.*, ICCSP, (Melmaruvathur, India), pp. 1514–1518, IEEE, Apr. 2014.

[9] S. Khanum and A. Firos, "A novel speaker identification system using feed forward neural networks," in *Proc. ICECDS*, ICECDS, (Chennai, India), pp. 3045–3047, IEEE, Aug. 2017.

[10] M. L. McLaren, M. I. Mandasari, and D. A. van Leeuwen, "Source normalization for language-independent speaker recognition using i-vectors," in *Proc. Odyssey*, (Singapore), pp. 55–61, Singapore:[Sn], June 2012.

[11] S. Tong, P. N. Garner, and H. Bourlard, "An Investigation of Deep Neural Networks for Multilingual Speech Recognition Training and Adaptation," in *Proc. Interspeech*, (Stockholm, Sweden), pp. 714–718, ISCA, Aug. 2017.

[12] W. Cai, J. Chen, and M. Li, "Exploring the Encoding Layer and Loss Function in End-to-End Speaker and Language Recognition System," in *Proc. Odyssey*, (Les Sables d'Olonne, France), pp. 74–81, ISCA, June 2018.

[13] D. Snyder, D. Garcia-Romero, G. ell, D. Povey, and S. Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition," in *Proc. ICASSP*, ICASSP, (Calgary, AB, Canada), pp. 5329–5333, IEEE, Apr. 2018.

[14] I. Shahin, "Speaker Identification in a Shouted Talking Environment Based on Novel Third-Order Circular Suprasegmental Hidden Markov Models," *Circuits, Systems, and Signal Processing*, vol. 35, pp. 3770–3792, Dec. 2015.

[15] I.-J. Ding and J.-Y. Shi, "Kinect microphone array-based speech and speaker recognition for the exhibition control of humanoid robots," *Comput. Electr. Eng.*, vol. 62, pp. 719–729, Aug. 2017.

[16] J. Kacur, "Modifications of KNN classifier for speaker identification system," in *Proc. of the International Symposium Electronics in Marine*, ELMAR, (Zadar, Croatia), pp. 35–38, IEEE, Sept. 2016.

[17] P. Bansal, S. A. Imam, and R. Bharti, "Speaker recognition using MFCC, shifted MFCC with vector quantization and fuzzy," in *Proc. ICSCTI*, ICSCTI, (Faridabad, India), pp. 41–44, IEEE, Oct. 2015.

[18] P. Bansal and S. A. Imam, "Performance of speaker recognition system using shifted mfcc, delta spectral cepstral coefficient (DSCC) and Fuzzy techniques," *Int. J. Eng. Technol.*, vol. 7, no. 2.8, pp. 278–283, 2018.

[19] S. Rathor and R. S. Jadon, "Text indpendent speaker recognition using wavelet cepstral coefficient and butter worth filter," in *Proc. 8th ICCCNT*, ICCCNT'08, (Delhi, India), pp. 1–5, IEEE, July 2017.

[20] M. Singh, R. Singh, and A. Ross, "A comprehensive overview of biometric fusion," *Inf Fusion*, vol. 52, pp. 187–205, Dec. 2019.

[21] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep Speaker Recognition," in *Proc. Interspeech*, (Hyderabad, India), pp. 1086–1090, ISCA, Sept. 2018.

[22] L. Li, D. Wang, A. Rozi, and T. F. Zheng, "Cross-lingual speaker verification with deep feature learning," in *Proc. Asia-Pacific Sig. Inform. Proces. Assoc. Annual Sum. Conf.*, APSIPA, (Kuala Lumpur, Malaysia), pp. 1040–1044, IEEE, Dec. 2017.

[23] L. Chen and C. Wu, "Crossed-Time Delay Neural Network for Speaker Recognition," May 2020.

[24] P. Matejka, O. Glembek, O. Novotny, O. Plchot, F. Grezl, L. Burget, and J. H. Cernocky, "Analysis of DNN approaches to speaker identification," in *Proc. ICASSP*, ICASSP, (Shanghai, China), pp. 5100–5104, IEEE, Mar. 2016.

[25] M. A. Przybocki, A. F. Martin, and A. N. Le, "NIST Speaker Recognition Evaluations Utilizing the Mixer Corpora—2004, 2005, 2006," *IEEE Trans Audio Speech Lang Process*, vol. 15, pp. 1951–1959, Sept. 2007.

[26] B. G. Nagaraja and H. S. Jayanna, "Mono and Cross lingual speaker identification with the constraint of limited data," in *Proc. Int. Conf. Pattern Recog. Informat. Med. Eng.*, PRIME, (Salem, Tamilnadu, India), pp. 439–443, IEEE, Mar. 2012.

[27] B. G. Nagaraja and H. S. Jayanna, "Efficient window for monolingual and crosslingual speaker identification using MFCC," in *Proc. Int. Conf. Advan. Comput. Commu. Syst.*, (Coimbatore, India), pp. 1–4, IEEE, Dec. 2013.

[28] E. Casanova, A. C. Junior, C. Shulby, H. Pereira da Silva, P. L. d. P. Filho, A. Ferreira Cordeiro, V. de Oliveira Guedes, and S. M. Aluisio, "Speech2Phone: A Multilingual and Text Independent Speaker Identification Model," Feb. 2020.

[29] T. A. de Lima and M. D. Costa-Abreu, "A survey on automatic speech recognition systems for Portuguese language and its variations," *Comput Speech Lang*, vol. 62, p. 101055, July 2020.

[30] Kaggle, "DARPA-TIMIT speech dataset," Dec. 2019. Accessed 17 December 2019.

[31] FalaBrasil, "LapsBenchmark 16k repository," Sept. 2018. Accessed 17 December 2019.

[32] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "AISHELL-1: An open-source Mandarin speech corpus and a speech recognition baseline," in *Proc. 20th Conf. Oriental Chap. Int. Coord. Commit. Speech Databases and Speech I/O Systems and Assessment*, 20, (Seoul, South Korea), pp. 1–5, IEEE, Nov. 2017.

[33] K. Jones, S. Strassel, K. Walker, D. Graff, and J. Wright, "Call My Net Corpus: A Multilingual Corpus for Evaluation of Speaker Recognition Technology.," in *Proc. Interspeech*, ISCA'17, (Toronto, ON, Canada), pp. 2621–2624, ISCA, Aug. 2017.

[34] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *J Mach Learn Res*, vol. 12, pp. 2825–2830, Feb. 2011.