Cross-Linguistic Speaker Profiling: Evaluating Monolingual and Multilingual Recognition through Machine Learning and Mel-Frequency Cepstral Coefficients

Mohamed Lichouri*, Rayane Fares Anis Embarek, Khaled Lounnas, Rachida Djeradi
LCPTS-USTHB University

DOI: 10.5281/zenodo.10968159

***Abstract*— Voice recognition has become increasingly integral in diverse applications, from virtual assistants to security systems. This paper presents a comprehensive study on the development and evaluation of an automated speaker profiling model for Arabic and English languages. Leveraging Support Vector Machine (SVM) and Gaussian Naive Bayes (GNB) classifiers on Mel-Frequency Cepstral Coefficients (MFCC) (Lounnas et al, 2018), our model demonstrates robust performance, achieving high precision, recall, and F1- scores across various speakers in a Multilingual corpus. The study highlights the superiority of SVM over GNB and provides insights into the model's strengths and limitations. The results open avenues for potential applications in voice-controlled systems and lay the groundwork for future research directions, emphasizing the importance of dataset expansion and audio quality improvement. This research contributes to the advancement of voice-related technologies, offering promising implications for real-world implementations.**

Keywords— *Speaker profiling, Support Vector Machine, Gaussian Naive Bayes, Mel-Frequency Cepstral Coefficients, Multilingual Corpus.*

1 INTRODUCTION

IN the contemporary landscape, user identification holds critical significance across diverse applications, ranging from mobile devices to banking systems (Kinnunen and Li, 2010). As the safeguarding of personal information becomes paramount, conventional verification methods, such as passwords and access cards, prove susceptible to various security threats. The evolution of technology has spurred the development of more secure authentication approaches.

In our interconnected world (Bisio et al., 2018), where approximately half of the global population seamlessly transitions between two languages, understanding the impact of linguistic diversity on computer vision and machine learning is crucial. This paper delves into the intricate questions surrounding the recent integration of cross-modal biometric (Hanifa et al., 2021) matching tasks in real-world scenarios, as depicted in Fig. 1:

Q1. How does language independence influence the association between voices and identities?

Q2. Can a speaker be reliably identified, regardless of the spoken language?

While prior research has unveiled a significant correlation between an individual's voice and identity, with studies garnering attention, none have delved into the consequences of multiple languages on this association. Commonly used datasets like VoxCeleb, FVCeleb, and FVMatching lack language-level annotations, limiting our understanding of the influence of multiple languages on connecting voices with identities.

To comprehensively address these questions, we introduce the Multilingual Voice Identity (MVI) dataset. This distinctive dataset comprises 3,000 audio recordings in two languages—Arabic, and English. The MVI dataset allows for a nuanced analysis of how multiple languages affect the association between voices and identities. For Q1, we propose a cross-modal verification approach, examining the effect of multiple languages on voice-identity association. Additionally, the audio component of the dataset, annotated with samples in three languages, lays the groundwork for addressing Q2 through speaker profiling baselines.

In summary, our contributions are as follows:

- Introduction of a cross-modal verification approach to analyze the impact of multiple languages on voice-identity association.
- Exploration of the multilingual challenge in speaker profiling.
- Presentation of the MVI dataset, encompassing 3,000 language-annotated audio clips, with 1,000 utterances per language, recorded by 10 speakers of the two gender.

These investigations carry practical implications, especially in applications such as secure voice authentication across diverse linguistic contexts. The subsequent sections are organized as follows: Section 2 reviews pertinent literature and existing datasets

*Mohamed LICHOURI e-mail: mlichouri@usthb.dz

related to the introduced questions. Section 3 details the nature of the proposed MVI dataset, while Sections 4 and 5 present experimental evidence addressing both questions. Finally, Section 6 concludes the paper.

2 RELATED WORKS

Speaker profiling (Schilling and Marsters, 2015), (Kalluri et al., 2020) is the process of estimating speaker characteristics, such as age, height, and other demographic information, from their speech data (Rajaa et al., 2021). It has a wide range of applications in various fields, including forensics, recommendation systems, and physical characteristics estimation. Speaker profiling can be carried out using different methods, such as:

- Aural-perceptual methods: These methods involve listening to the speech sample and identifying the speaker's characteristics based on the impression given by the listener (Kalluri et al., 2021).
- Acoustic phonetic analysis (Guille´n-Nieto and Stein, 2022): This method focuses on the analysis of phonetic features (Kulshreshtha et al., 2012) in the speech sample to derive information about the speaker.
- Automated methods: These methods use automatic speaker profiling technology and algorithms to process and analyze speech data, resulting in the estimation of speaker characteristics.

Some of the speaker characteristics that can be estimated through speaker profiling include gender, age, sociolect (profession, education level), foreign accent, native dialect, and medical conditions (Kalluri et al., 2021). The accuracy of speaker profiling depends on the method used and the complexity of the speech data. However, it is essential to note that there are complications in identifying each of the various speaker attributes, and the focus on different methods may vary depending on the specific application and context.

Previous approaches to speaker profiling in Arabic, French, and English have utilized various techniques, including automatic speech recognition, cross-language acoustic models, and voice biometrics distinction. For instance, a contrastive study of Arabic and English for improved language teaching and speech processing has been conducted using neural networks (Dib, 2018). Additionally, the use of common acoustic models for Arabic and English speech recognition systems has been evaluated (Alotaibi et al., 2012). Furthermore, voice biometrics distinction between English, French, Arabic, and Spanish has been explored using sound cleaner filtering and SpeechPro SIS II Anal (Akhdar and Jasra, 2020). These approaches demonstrate the diverse methods employed to address speaker profiling in multilingual contexts.

3 DATA COLLECTION AND PREPARATION

The dataset employed for speaker profiling is derived from vocal recordings of instructions captured using a mobile phone, specifically named the Multilingual Voice Identity (MVI) dataset. To enhance the granularity of our analysis, each instruction was meticulously segmented into 10 distinct parts using Praat software. This segmentation process contributes to a comprehensive exploration of the intricate vocal characteristics exhibited by the speakers within the MVI dataset, ensuring a nuanced understanding of the diverse and unique elements embedded in the recorded instructions.

A. Data Collection

Vocal recordings were conducted in diverse conditions, including indoor and outdoor environments, to capture the diversity of situations in which speakers might use their voices to give instructions. Despite the technical limitations of a phone, efforts were made to obtain the best possible audio quality.

The data used for speaker profiling come from a diverse group of speakers, representing different age groups, genders, native languages, and dialects. This diversity establishes a rich and representative vocal database, facilitating accurate and reliable speaker profiling.

For the collection of this data, we used the mobile application waveEditor (Lounnas et al., 2020), (Lounnas et al., 2022), an audio software for recording, editing, and manipulating audio files. The vocal segments obtained through Praat target specific aspects of the speakers' voices, providing additional granularity when analyzing vocal characteristics such as timbre, rhythm, emphasis, and other distinctive traits unique to each speaker.

We collected vocal recordings from the speakers using a set of 10 specific commands in English, French, and Arabic. The commands, listed below, were carefully chosen to control an engine and were recorded in various contexts to reflect realistic situations. Additionally, to ensure linguistic diversity and cultural relevance, we translated these 10 commands into French and English. Subsequently, a completely new corpus was recorded, featuring distinct vocalizations for the translated commands. It is important to note that the original Arabic commands, as detailed in (Lichouri et al., 2023), served as the foundation for this translation process, maintaining consistency while expanding the dataset.

English Commands: 1- forward 2- stop 3- start the engine 4- go 5- left 6- right 7- reduce speed 8- speed up 9- turn off the engine 10- to the back

Table I presents some statistics on the profiles of the speakers chosen for the database recording.

TABLE I
SPEAKER CHARACTERISTICS

Identifier	Age	Gender	City	Dialect	English Level
Ines	19	F	Algiers	Algiers dialect	B2
Maria	20	F	Algiers	Algiers dialect	C1
Mariah	20	F	Algiers	Algiers dialect	B2
Nadir	20	M	Algiers	Algiers dialect	B2
Nour	20	F	Algiers	Algiers dialect	C1
Rayane	22	M	Algiers	Algiers dialect	C1
Sarah	22	F	Algiers	Algiers dialect	C1
Yacine	27	M	Algiers	Algiers dialect	B2
Khaled	30	F	Algiers	Algiers dialect	B1
Mohamed	38	M	Blida	Blida dialect	B2

B. Analysis of Speaker Characteristics

The statistics presented in Table I offer insights into the diverse group of speakers whose vocal data were utilized for the profiling task. Here are some key observations:

- **Age Diversity:** Unfortunately, the current dataset is limited to speakers aged 19-38. While the initial age range description (19-38) might suggest a broader spectrum, as highlighted in Table I, the first speakers fall within a narrower range (19-22). This limits the generalizability of our findings regarding the impact of age on speaker profiing. Future research with a more diverse age range encompassing childhood, adolescence, youth, adulthood, and elderhood is necessary to provide a more complete picture of how language influences speaker profiing across different age groups.
- **Gender Representation:** The dataset includes both male and female speakers, contributing to a balanced representation. This gender diversity is essential for creating a comprehensive vocal database that accommodates the characteristics of both genders.
- **City and Dialect Variability:** While the current dataset incorporates speakers from two Algerian cities (Algiers and Blida), representing a limited geographic range, these diversities enrich the corpus by accounting for variations in pronunciation and linguistic nuances specific to these regions. However, a broader dialectal scope encompassing a larger set of Algerian cities would be necessary to achieve a more comprehensive analysis of dialectal variability's influence on speaker profiing systems. Future research efforts will aim to incorporate speakers from a wider range of Algerian regions to strengthen the generalizability of our findings.
- **Language Proficiency:** The English proficiency levels of speakers range from B1 to C1, covering a spectrum from intermediate to advanced proficiency. This language variation allows for the exploration of how different language skills might influence speaker profiling.

In summary, the speaker characteristics demonstrate a deliberate effort to create a diverse dataset that encompasses various demographic factors. This diversity enhances the dataset's representativeness and ensures a more robust analysis of speaker profiling across different age groups, genders, cities, dialects, and language proficiency levels. We have thus created a diverse and multilingual vocal corpus, providing a robust foundation for speaker profiling analysis, irrespective of the language.

4 EXPERIMENTAL SETUP

In the context of the conducted experiments, a vocal corpus comprising samples from 10 speakers for each language was utilized. Speakers were deliberately chosen to encompass a diversity of age, gender, dialect or accent, and English proficiency levels. To thoroughly investigate the impact of language on the performance of the speaker profiling system, three experiments were designed based on the spoken language: Arabic, English, and Multilingual (Arabic and English). We will assess the system's performance in terms of recognizing gender, age, accent, dialect, and geographical region of the speakers.

The evaluation will encompass an analysis of system performance using metrics such as precision, recall, and F-measure for each identification task. This comprehensive approach allows us to delve into the effectiveness of the system across various linguistic and demographic dimensions, providing valuable insights into its capabilities and potential limitations.

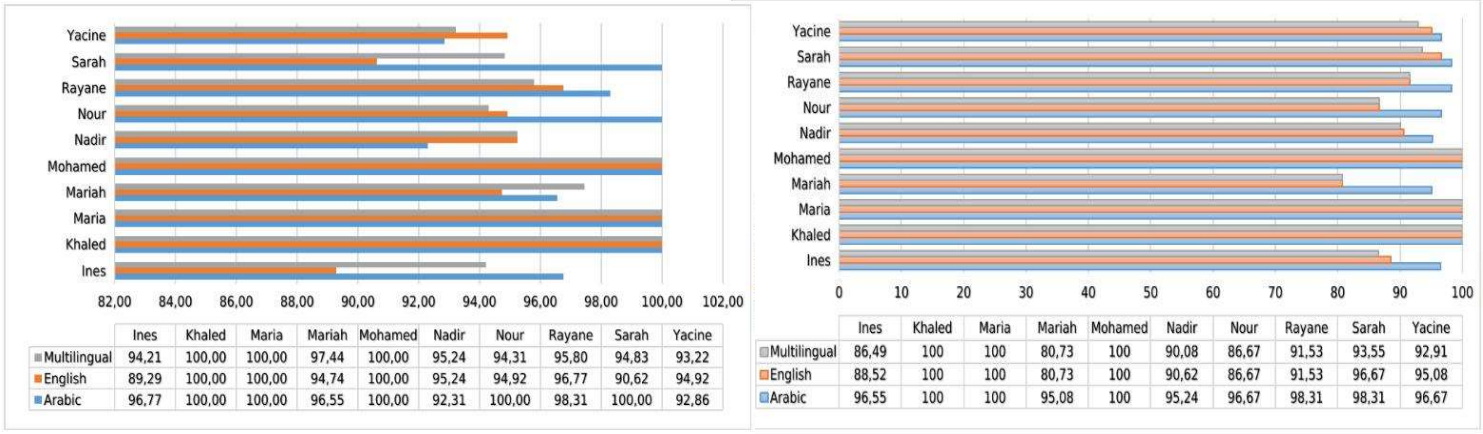
5 RESULTS ANALYSIS

In the context of the conducted experiments, a vocal corpus comprising samples from 10 speakers for each language was used. Speakers were selected to represent diversity in age, gender, dialect or accent, and English proficiency. To thoroughly investigate the impact of language on the performance of the speaker profiling system, three experiments were conducted for Arabic, English, and Multilingual (Arabic and English) languages.

This summary table (see Table II) summarizes the macro-average performance of the SVM classifier, providing a concise overview of its effectiveness across various language corpora. The SVM classifier achieves the highest F1-score of 97.68% in speaker profiling within the Arabic corpus. However, a slight decrease in performance is observed for the English and Multilingual corpora, with approximately -2% and -1%, respectively

TABLE II
SUMMARY OF RESULTS (MACRO) FOR AUTOMATED SPEAKER PROFILING USING THE SVM CLASSIFIER: ARABIC, ENGLISH, AND MULTILINGUAL CORPORA.

Corpus	Precision	Recall	F1-score
Arabic	0.9795	0.9767	**0.9768**
English	0.9592	0.9567	**0.9565**
Multilingual	0.9661	0.9650	**0.9650**



	Ines	Khaled	Maria	Mariah	Mohamed	Nadir	Nour	Rayane	Sarah	Yacine
Multilingual	94,21	100,00	100,00	97,44	100,00	95,24	94,31	95,80	94,83	93,22
English	89,29	100,00	100,00	94,74	100,00	95,24	94,92	96,77	90,62	94,92
Arabic	96,77	100,00	100,00	96,55	100,00	92,31	100,00	98,31	100,00	92,86

	Ines	Khaled	Maria	Mariah	Mohamed	Nadir	Nour	Rayane	Sarah	Yacine
Multilingual	86,49	100	100	80,73	100	90,08	86,67	91,53	93,55	92,91
English	88,52	100	100	80,73	100	90,62	86,67	91,53	96,67	95,08
Arabic	96,55	100	100	95,08	100	95,24	96,67	98,31	98,31	96,67

Fig. 1. Individual Speaker Proficiency: Obtained F1-Scores by SVM (above) and GNB (bellow) Classifiers Across Different Language Corpora

This summary table (see Table III) outlines the macro-average performance of the GNB classifier, presenting an overview of its effectiveness across distinct language corpora. The GNB classifier demonstrates the highest F1-score of 97.68% in speaker profiling within the Arabic corpus. However, a decrease in performance is noticeable for the English and Multilingual corpora, with approximately -4% and -5%, respectively. Despite these reductions, the GNB classifier maintains strong precision and recall values across all corpora.

TABLE III
SUMMARY OF RESULTS (MACRO) FOR AUTOMATED SPEAKER PROFILING USING THE GNB CLASSIFIER: ARABIC, ENGLISH, AND MULTILINGUAL CORPORA.

Corpus	Precision	Recall	F1-score
Arabic	0.9778	0.9767	**0.9768**
English	0.9376	0.9367	**0.9363**
Multilingual	0.9255	0.9233	**0.9220**

6 DISCUSSION OF PERFORMANCES

The obtained F1-scores for each speaker in the Arabic, English, and Multilingual corpora, as analyzed by both SVM and GNB classifiers, reveal valuable insights into the performance of our automated speaker profiling system.

The SVM classifier demonstrates consistently high precision, recall, and F1-scores, with the Arabic corpus exhibiting the highest overall performance. Specifically, the Arabic model achieves an outstanding F1-score of 97.68

In contrast, the GNB classifier also showcases strong performance across the three corpora. The Arabic corpus once again leads with a remarkable F1-score of 97.68

The general trend of SVM outperforming GNB is consistent with expectations, given SVM's resilience to noise and superior generalization abilities, especially with smaller datasets. However, a nuanced examination of the confusion matrix highlights that recordings without background noise are rarely confused with other speakers, suggesting potential areas for model optimization, particularly in handling noise.

Exploring the differences between the Multilingual, Arabic, and English models, the SVM results indicate that the Arabic model excels, followed by the Multilingual model and then the English model. Similarly, the GNB results underscore the dominance of the Arabic model, followed by the English model and the Multilingual model.

These findings provide valuable insights into the strengths and areas for improvement of our speaker profiling model. The exceptional performance of the Arabic model suggests its robustness in handling diverse linguistic nuances, while the observed trends between SVM and GNB offer valuable considerations for optimizing the model further. This experiment serves as a pivotal step towards comprehensive speaker profiling, opening avenues for future research in refining our understanding of automated speaker profiling.

A. Limitations and Future Perspectives

In the scope of this study, we face certain limitations that constrain our results. Firstly, the limited size of the corpus poses a major challenge. It is well-known that the performance and generalization capability of a model improve with the size of the training set. In our case, the corpus consists of only 10 speakers providing 10 instructions in English and 10 in Arabic. Therefore, this study should be considered preliminary, but it offers interesting avenues for future research using larger datasets.

Another limitation concerns the quality of audio recordings. We used different mobile phones, specific to each speaker, to perform the recordings. This approach may introduce undesirable noise into the model, thus altering fidelity to the original voice.

In future studies, it would be preferable to use dedicated professional microphones to record the corpus. This would capture voice subtleties, providing the model with better audio analysis capabilities with more details. We observed that sometimes the models struggle to distinguish between male and female voices, which could be due to the limited corpus size or audio quality making the task of distinction more complex for the model.

Regarding future perspectives, in addition to improving audio quality, it would be interesting to explore other learning models to determine which is best suited for our application. The obtained results could also be used to further optimize existing models. Since machine learning is not an exact science, multiple trials will be necessary to assess the effectiveness of modifications made to learning algorithms.

It is important to consider these limitations when interpreting the results and regard them as pathways for future research aimed at overcoming these challenges and deepening our understanding of speaker profiling.

7 CONCLUSION

In conclusion, our study successfully developed and evaluated an automated speaker profiling model for Arabic and English languages. The results demonstrate exceptional performance, particularly with the SVM classifier achieving remarkable accuracy: 97.68% for Arabic, 95.65% for English, and 96.50% for the multilingual corpus. These findings highlight the model's effectiveness in speaker identification, even when speakers are not native English speakers. The strong performance with Arabic, the first language for all participants, suggests the model's ability to capture speaker-specific characteristics that transcend language proficiency.

It's noteworthy that the multilingual corpus performance (96.50%) falls slightly between the individual language results. This could be attributed to the influence of varying English proficiency levels among speakers. Future research could explore strategies to optimize the model's performance for speakers with diverse language backgrounds.

These findings not only affirm the accuracy of the developed model in identifying speakers based on language but also showcase its potential applications in voice-controlled systems, personalized user interfaces, and security systems with precise voice recognition.

Despite the overall success, it is crucial to acknowledge the study's limitations. The small corpus size and diverse recording devices may have introduced some noise into the model. As a valuable preliminary investigation, this study sets the stage for future research endeavors.

REFERENCES

(Akhdar and Jasra, 2020) Akhdar, S. and Jasra, S. K. (2020). Voice biometrics distinction between english, french, arabic and spanish using sound cleaner filtering and speechpro sis ii analysis for same individual identification in multilingual societies. *Journal of Emerging Forensic Sciences Research*, 5(1):65–72.

(Alotaibi et al., 2012) Alotaibi, Y. A., Selouani, S.-A., Alghamdi, M. M., and Meftah, A. H. (2012). Arabic and english speech recognition using cross-language acoustic models. In *2012 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA)*, pages 40–44. IEEE.

(Bisio et al., 2018) Bisio, I., Garibotto, C., Grattarola, A., Lavagetto, F., and Sciarrone, A. (2018). Smart and robust speaker recognition for context-aware in-vehicle applications. *IEEE Transactions on Vehicular Technology*, 67(9):8808–8821.

(Dib, 2018) Dib, M. (2018). *Automatic Speech Recognition of Arabic Phonemes with Neural Networks: A Contrastive Study of Arabic and English*. Springer.

(Guille´n-Nieto and Stein, 2022) Guille´n-Nieto, V. and Stein, D. (2022). *Language as evidence: Doing forensic linguistics*. Springer Nature.

(Hanifa et al., 2021) Hanifa, R. M., Isa, K., and Mohamad, S. (2021). A review on speaker recognition: Technology and challenges. *Computers & Electrical Engineering*, 90:107005.

(Kalluri et al., 2020) Kalluri, S. B., Vijayasenan, D., and Ganapathy, S. (2020). Automatic speaker profiling from short duration speech data. *Speech Communication*, 121:16–28.

(Kalluri et al., 2021) Kalluri, S. B., Vijayasenan, D., Ganapathy, S., Krishnan, P., et al. (2021). Nisp: a multi-lingual multi-accent dataset for speaker profiling. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6953–6957. IEEE.

(Kinnunen and Li, 2010) Kinnunen, T. and Li, H. (2010). An overview of text-independent speaker recognition: From features to supervectors. *Speech communication*, 52(1):12–40.

(Kulshreshtha et al., 2012) Kulshreshtha, M., Singh, C., and Sharma, R. (2012). Speaker profiling: The study of acoustic characteristics based on phonetic features of hindi dialects for forensic speaker identification. *Forensic speaker recognition: Law enforcement and counter-terrorism*, pages 71–100.

(Lichouri et al., 2023) Lichouri, M., Lounnas, K., and Bakri, A. (2023). Toward building another arabic voice command dataset for multiple speech processing tasks. In *2023 International Conference on Advances in Electronics, Control and Communication Systems (ICAECCS)*, pages 1–5. IEEE.

(Lounnas et al., 2022) Lounnas, K., Abbas, M., Lichouri, M., Hamidi, M., Satori, H., and Teffahi, H. (2022). Enhancement of spoken digits recognition for under-resourced languages: case of algerian and moroccan dialects. *International Journal of Speech Technology*, 25(2):443–455.

(Lounnas et al., 2018) Lounnas, K., Demri, L., Falek, L., and Teffahi, H. (2018). Automatic language identification for berber and arabic languages using prosodic features. In *2018 International Conference on Electrical Sciences and Technologies in Maghreb (CISTEM)*, pages 1–4. IEEE.

(Lounnas et al., 2020) Lounnas, K., Satori, H., Hamidi, M., Teffahi, H., Abbas, M., and Lichouri, M. (2020). Cliasr: a combined automatic speech recognition and language identification system. In *2020 1st international conference on innovative research in applied science, engineering and Technology (IRASET)*, pages 1–5. IEEE.

(Rajaa et al., 2021) Rajaa, S., Van Tung, P., and Siong, C. E. (2021). Learning speaker representation with semi-supervised learning approach for speaker profiling. *arXiv preprint arXiv:2110.13653*.

(Schilling and Marsters, 2015) Schilling, N. and Marsters, A. (2015). Unmasking identity: Speaker profiling for forensic linguistic purposes. *Annual Review of Applied Linguistics*, 35:195–214.