Analyzing the Impact of Synthetic Speech on Spoken Language Identification

Khaled Lounnas*
Speech Communication and Signal Processing Laboratory,
Faculty of Electrical Engineering
USTHB, Algiers, Algeria
CRSTDLA, Algiers,Algeria
Email: k.lounnas@crstdla.dz

Abderrahmane Gamgani
Algiers01 University, Algiers,Algeria
a.gamgani@univ-alger.dz

Imad Feth-Ennour Aliane
Algiers01 University, Algiers,Algeria
aliane.imad@hotmail.com

Received: 29/11/2024 **Accepted:** 13/12/2024 **Published:** 30/12/2024

Abstract:

This research explores how synthetic speech influences the performance of spoken language identification systems by examining various feature types (acoustic, temporal, and rhythmic) across multiple machine learning and deep learning architectures. The study utilized Multi-Layer Perceptron (MLP), Support Vector Machines (SVM), and Long Short-Term Memory networks (LSTM) to evaluate three distinct scenarios: identifying languages from natural speech, synthetic speech, and a mix of both. Furthermore, the study investigated whether integrating all feature types could enhance system performance. The results revealed that the Mel spectrogram consistently emerged as the most effective feature across all tested models, with MLP and LSTM achieving the best overall results. In fact, the Mel spectrogram attained a remarkable accuracy rate of 100%, establishing itself as the top-performing feature. Similarly, MFCC also reached 100% accuracy in the synthetic speech scenario, highlighting its strength as the second most effective feature. Notably, combining all features did not always lead to performance improvements, underscoring the importance of strategic feature selection. The study also tackled challenges like variability in natural speech recordings and imbalances in dataset distribution, emphasizing the necessity of robust data augmentation methods. By shedding light on the interactions between feature types, model

architectures, and speech data sources, this research advances the development of more accurate and resilient spoken language identification systems.

Keywords: Spoken Language Identification, Arabic language, Data Augmentation, Synthetic Corpus, LSTM, MLP.

* Corresponding author: Khaled Lounnas

تحليل تأثير الكلام الاصطناعي على تحديد اللغة المنطوقة

الملخص

اتستكشف هذه الدراسة تأثير الكلام الاصطناعي على أداء أنظمة تحديد اللغة المنطوقة من خلال تحليل أنواع مختلفة من الميزات – الصوتية والزمنية والإيقاعية – عبر عدة هياكل من التعلم الآلي والتعلم العميق. استخدمت الدراسة شبكات البرسيبترون (Perceptron) متعدد الطبقات(MLP) ، وآلات المتجهات الداعمة (SVM)، وشبكات الذاكرة طويلة وقصيرة المدى (LSTM) لتقييم ثلاثة سيناريوهات مختلفة: تحديد اللغات من الكلام الطبيعي، والكلام الاصطناعي، ومزيج من الاثنين. علاوة على ذلك، بحثت الدراسة في ما إذا كان دمج جميع أنواع الميزات يمكن أن يعزز أداء النظام. كشفت النتائج أن طيف ميل الصوتي كان باستمرار الميزة الأكثر فعالية عبر جميع النماذج التي تم اختبارها، حيث حقق كل من MLP و LSTM أفضل النتائج الإجمالية. في الواقع، حقق طيف ميل دقة مذهلة بنسبة 100٪، مما يجعله أفضل ميزة أداء. وبالمثل، حققت ميزات MFCC أيضًا دقة بنسبة 100٪ في سيناريو الكلام الاصطناعي، مما يبرز قوتها كثاني أفضل ميزة. ومن الجدير بالذكر أن الجمع بين جميع الميزات لم يؤدِ دائمًا إلى تحسين الأداء، مما يؤكد أهمية الاختيار الاستراتيجي للميزات. كما تناولت الدراسة تحديات مثل التباين في تسجيلات الكلام الطبيعي وعدم التوازن في توزيع البيانات، مشددةً على ضرورة وجود طرق قوية لتعزيز البيانات. ومن خلال تسليط الضوء على التفاعلات بين أنواع الميزات، وهياكل النماذج، ومصادر بيانات الكلام، تعزز هذه الدراسة تطوير أنظمة تحديد اللغة المنطوقة بدقة ومرونة أكبر.

الكلمات المفتاحية: تحديد اللغة المنطوقة، اللغة العربية، زيادة البيانات، مجموعة البيانات الاصطناعية، LSTM، MLP.

Analyse de l'impact de la parole synthétique sur l'identification de la langue parlée

Résumé

Cette recherche examine l'impact de la parole synthétique sur les performances des systèmes d'identification des langues parlées en analysant différents types de caractéristiques (acoustiques, temporelles et rythmiques) à travers plusieurs architectures d'apprentissage automatique et profond. Nous avons utilisé les réseaux de Perceptron Multicouche (MLP), les Machines à Vecteurs de Support (SVM) et les réseaux Long Short-Term Memory (LSTM) pour évaluer l'identification des langues à partir de trois scénarios distincts : la parole naturelle, la parole synthétique et un mélange des deux. En outre, nous avons exploré si l'intégration de tous les types de caractéristiques pourrait améliorer les performances du système. Les résultats ont révélé que le spectrogramme de Mel était systématiquement la caractéristique la plus efficace pour tous les modèles testés, avec les MLP et LSTM obtenant les meilleurs résultats globaux. En effet, le spectrogramme de Mel a atteint un taux d'exactitude remarquable de 100 %, se positionnant comme la caractéristique la plus performante. De même, les coefficients cepstraux en fréquences de Mel (MFCC) ont également atteint une précision de 100 % dans le scénario de la parole synthétique, confirmant leur efficacité en tant que deuxième meilleure caractéristique. Il est intéressant de noter que la combinaison de toutes les caractéristiques n'a pas toujours amélioré les performances, soulignant l'importance d'une sélection stratégique des caractéristiques. L'étude a également abordé des défis tels que la variabilité des enregistrements de parole naturelle et les déséquilibres dans la distribution des ensembles de données, mettant en avant la nécessité de méthodes robustes d'augmentation des données. En éclairant les interactions entre les types de caractéristiques, les architectures de modèles et les sources de données vocales, cette recherche contribue au développement de systèmes d'identification des langues parlées plus précis et plus résilients.

Mots clés : Identification de la langue parlée, langue arabe, augmentation des données, corpus synthétique, LSTM, LMP.

INTRODUCTION

The rapid development of speech synthesis technology has resulted in synthetic speech that closely resembles natural human speech. This progress brings both opportunities and challenges to spoken language identification systems, which are crucial in applications such as multilingual customer support, voice-activated assistants, and language learning platforms. While synthetic speech can enrich training datasets and enhance model robustness, it also introduces distinct characteristics that might affect system performance. This research explores the impact of synthetic speech on spoken language identification by assessing a variety of feature sets and machine learning models, including a deep learning approach. Three scenarios are examined: natural speech, synthetic speech, and a combination of both. The study utilizes acoustic, temporal, and rhythmic features, evaluating the performance of three different models—Multi-Layer Perceptron (MLP), Support Vector Machines (SVM), and Long Short Term Memory (LSTM) networks.

The goal is to understand how synthetic speech influences the accuracy and reliability of language identification systems. By investigating how different features and models handle the variability between synthetic and natural speech, this work aims to provide valuable insights for improving spoken language identification systems under diverse and real-world conditions. The results could have practical applications in enhancing multilingual customer support, refining voice controlled technologies, and advancing adaptive language learning systems.

1. RELATED WORK

The study of language identification has made great strides over the years, aiming to improve accuracy and reliability. In this section, we review important research on speech synthesis, language identification, and how synthetic voices affect language processing systems. These studies show recent advancements and highlight the gaps that our research aims to fill.

A. Traditional Approaches to Language Identification

Early language identification systems relied on statistical models and acoustic feature extraction from human speech. Two common techniques used were:

- *Gaussian Mixture Models (GMMs):* GMMs were used to classify languages based on the spectral features of speech signals (Kumar et al., 2010; Wicaksana et al., 2021). These models represented the distribution of features with a

mixture of Gaussian distributions, offering a probabilistic approach to language classification.

- *Hidden Markov Models (HMMs):* HMMs utilized sequences of phonetic units to identify languages. By modeling transitions between phonetic states, HMMs captured the linguistic patterns unique to different languages (Kumar et al., 2005; Sefara et al., 2019), allowing for effective classification.

B. Deep Learning Techniques in Language Identification

The rise of deep learning (Singh et al., 2021) has transformed language identification by enabling models to learn complex patterns directly from raw audio data:

- *Convolutional Neural Networks (CNNs):* CNNs are used to capture the local temporal and spectral features of speech. Their ability to learn hierarchical patterns from spectrograms makes them highly effective for language identification tasks (Ganapathy et al., 2014; Singh et al., 2021).
- *Recurrent Neural Networks (RNNs):* RNNs (Alashban et al., 2022), including Long Short-Term Memory (LSTM) networks (Zazo et al., 2016; Gelly et al., 2017), improve the modeling of sequential data in speech. They can capture long-term dependencies in audio signals, making them well-suited for identifying languages from continuous speech.
- *Feature Extraction Methods:* Techniques such as Mel Frequency Cepstral Coefficients (MFCCs) and Mel spectrograms remain important. MFCCs (Biswaset al., 2023) provide a compact representation of speech power, while spectrograms offer a visual depiction of frequency content over time. Both are effective inputs for deep learning models.

C. Impact of Synthetic Voices on Language Identification

The growing use of synthetic voices in various applications highlights the need to understand their effects on language identification systems: Research shows that synthetic speech (Duffy et al.,1992) can be more difficult to understand than natural speech. Synthetic voices often cause delays in recognizing words because the phonetic processor struggles to extract phonemes. This delay can affect comprehension, as more cognitive effort is needed to decode the speech rather than to grasp its meaning.

Recent studies indicate that deep learning models are increasingly effective in language identification, even with synthetic speech data (Ambili & Roy, 2023).

These models can achieve high accuracy in identifying languages from audio clips, showing promise for enhanced language processing capabilities in synthetic contexts.

D. Data Augmentation Techniques

Data augmentation has proven to be a crucial strategy for enhancing the generalization and robustness of language identification systems:

• *Common Techniques:* Methods such as adding noise, pitch alteration, and time-stretching (Maguolo et al., 2021) are used to simulate various real-world conditions. These techniques increase the diversity of the training data, helping models become more resilient to variations in speech input.

• Ambili and Roy demonstrated significant improvements in Indian language identification systems through data augmentation (Ambili & Roy, 2023). Their study addressed the challenge of phonetic similarity in spoken language identification by proposing a synthetic voice data augmentation method. Using pre-trained models (VGG16, RESNET50, Inception-v3) and various classifiers, the research found that synthetic audio samples improved accuracy by 17%. Building upon these foundational studies, our research aims to provide a comprehensive understanding of the impact of synthetic voices on language identification systems. We address this by:

Corpus Design: We have created a unique corpus that includes both normal and synthetic voices for three languages: Arabic, English, and French. This dataset serves as a comprehensive resource for evaluating the performance of language identification systems under varying voice conditions.

• *Data Augmentation*: We employ a range of data augmentation techniques to extend the corpus. These methods simulate diverse acoustic scenarios and enhance the robustness of our models.

• *Evaluation:* Our study systematically evaluates the impact of voice type on language identification performance. By analyzing performance metrics, we aim to identify any biases or limitations introduced by synthetic voices and propose strategies to mitigate these effects.

Our research focuses on identifying the Arabic language among English and French using both machine and deep learning technique that is infrequently applied to synthetic voice-based speech commands. Identifying a language in very short utterances presents significant challenges due to the limited linguistic information available, which complicates the task compared to sentence-level language identification (LID). This focus on multilingual (based vocal command) LID represents a more intricate challenge. Through

this study, we aim to enhance the understanding of how synthetic voices impact language identification systems and contribute to the development of more adaptable and robust models.

2. PROPOSED SPOKEN LANGUAGE IDENTIFICATION FRAMEWORK

The motivation and framework for spoken language identification are illustrated in Figure 1. The figure shows a process involving both synthetic and natural speech. Synthetic speech is generated, and natural speech is recorded, with both types being modified by adding noise and applying roll data augmentation. The modified datasets are then combined, and different features set are extracted. The resulted features are split into training and testing sets. A model is trained using the training data and then used to identify languages (Arabic, French, English) in the testing set. The accuracy of various machine learning models and deep learning method is compared to assess their performance

A. Motivation

This research looks at how synthetic voices affect language identification systems, especially as synthetic voices become more common in tools like virtual assistants and language learning apps. The goal is to make these systems more accurate and reliable by tackling the unique challenges of synthetic speech, which is different from natural human speech in sound and pronunciation. By finding the best feature settings, we can improve how well these systems work with different types of speech. In the end, this study aims to help create technology that works well for everyone, making sure language identification is accurate for all users.

B. Proposed Spoken Identification Framework

The proposed scheme, illustrated in Figure 1, improves spoken language identification systems by integrating synthetic and natural speech recordings. These recordings are enhanced using data augmentation techniques like adding noise and applying roll algorithms. The augmented data is then combined, and features related to sound and spectrum are extracted for training the models. Both machine learning and deep learning algorithms use these features to create an acoustic/spectral model. This model is tested to classify speech into

languages such as Arabic, French, and English, and the performance of the different machine learning and deep learning approaches is compared.

C. Preprocessing

In the preprocessing phase, data augmentation helps to address class imbalance issues, reduces overfitting, and serves as a regularizer during model training. Techniques like adding noise and applying roll transformations increase the data size by modifying existing samples. These variations, such as noise and time shifts, improve the robustness of machine learning and LSTM deep learning models. By working with a larger, more varied dataset, these methods enhance the models' performance, leading to better prediction results.

D. Description of Features

In our study, we utilized a framework based on Librosa (Mc Fee et al., 2015), which incorporates a variety of spectral features and rhythm characteristics. This framework enabled us to extract a comprehensive set of features to enhance spoken language identification. The features used in this framework include:

1) MFCC coefficients (40): Mel-frequency cepstral coefficients (MFCCs) are widely used in speech and audio processing. They capture the short-term power spectrum of a sound and are highly effective in representing the phonetic content of speech.

2) Mel spectrogram (128) & Chroma Vector (12): The Mel spectrogram represents the power spectral density of a signal on a Mel scale of frequency, providing a detailed time-frequency representation of the audio.

Chroma vectors capture the 12 different pitch classes, highlighting the harmonic and melodic content of the audio signal.

3) Spectral contrast (7) & Tonnetz (6): Spectral contrast measures the difference in amplitude between peaks and valleys in a sound spectrum, offering insights into the timbral texture of the audio. The Tonnetz feature, based on the tonal centroid features, represents the harmonic relations in music, which can be useful in identifying tonal qualities in speech. In total, these features amount to 193 components, providing a rich and diverse set of data for improving the accuracy and robustness of spoken language identification models.

E. Model Description

In this section, we describe the framework used for all models in the experiment, with a specific focus on the LSTM based model configuration, Multi-Layer Perceptron (MLP), and Support Vector Machine (SVM) configurations:

- ***LSTM Layer***

The model begins with an LSTM layer consisting of 256 units. This layer is responsible for capturing temporal dependencies in the data, which is crucial for spoken language identification. The input shape is configured to handle one time step with a feature size matching the number of input features.

• *Dropout Layer:* Following the LSTM layer, a dropout layer with a dropout rate of 0.6 is applied. This layer helps prevent overfitting by randomly setting a fraction of the input units to zero during training, promoting the robustness of the model.

• *Dense Layer:* The final layer is a dense (fully connected) layer. This layer serves as the output layer of the model and uses a softmax activation function. The number of units in this layer corresponds to the number of classes in the label encoder, allowing the model to output a probability distribution over the possible spoken languages.

- ***Multi-Layer Perceptron (MLP)***

The MLP model used in the experiment consists of 1 to 3 dense layers. This architecture is designed to capture non-linear relationships in the data and deliver robust classification performance.

TABLE 1. HYPERPARAMETER GRID FOR NEURAL NETWORK FINE-TUNING

Hyperparameter	Values
Hidden Layer Sizes	{(100,), (50, 50), (50, 25, 10)}
Activation Functions	{relu, logistic, tanh}
Solvers	{adam, sgd, lbfgs}
Alpha loguniform	(1e-5, 1e-1)
Learning Rate	{constant, adaptive, nvscaling}
Initial Learning Rate	loguniform(1e-4, 1e-1)
Batch Size	{32, 64, 128}
Momentum	{0.9, 0.95, 0.99}

To improve the performance of the MLP classifier, we employed a randomized search approach for hyper parameter optimization. This technique systematically examines a variety of hyper parameter settings to identify the optimal combination that enhances both accuracy and generalization. The parameters included in our search grid are outlined in the aforementioned table (see Table 1).

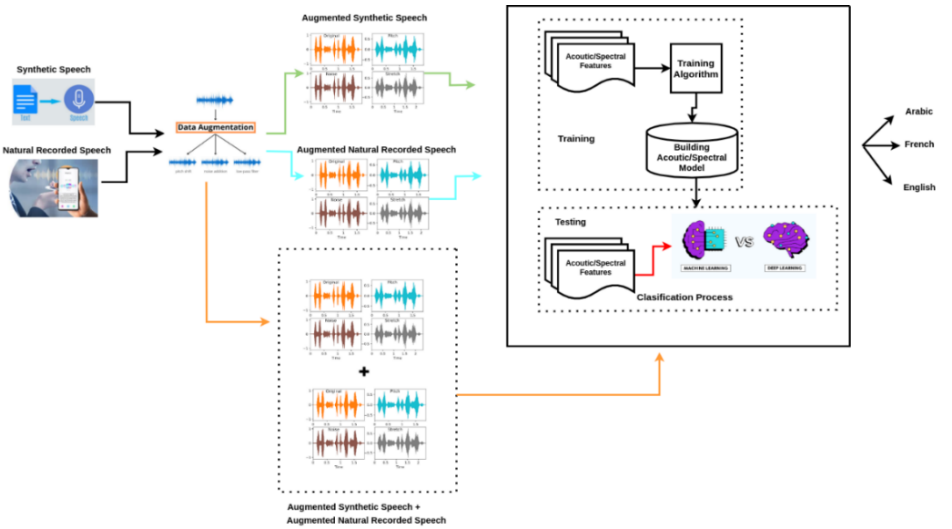

Figure 1. The Proposed Approach for Arabic Spoken command Recognition

- Support Vector Machine (SVM)

For the SVM model, we use the default configuration provided by the Scikit learn library. This includes a radial basis function (RBF) kernel with default values for the regularization parameter *C* and gamma. The default settings are used to provide a baseline performance for comparison with other models. The described frameworks ensure that the models are capable of learning complex patterns and handling various aspects of spoken language identification. The LSTM-based model focuses on capturing temporal dependencies, the MLP is designed to learn non-linear relationships, and the SVM provides a baseline for comparison with its default configuration. This comprehensive approach allows for effective evaluation of different modeling strategies in the context of spoken language identification tasks.

3. EXPERIMENTS AND RESULTS

The following section presents the results and discusses the effectiveness of various techniques. It includes a detailed overview of the datasets used, including their statistical characteristics, and an explanation of the different data augmentation techniques employed. These details aim to offer a thorough understanding of the impact and performance of the proposed methods.

A. Corpus

The corpus for our language identification task contains over 11,000 speech samples in Arabic, French, and English (as depicted in Table 2). This dataset includes two types of recordings: synthetic (computer-generated) and normal (recorded from real people).

TABLE 2. DETAILS ON THE DEVELOPPED CORPUS (NORMAL & SYNTHETIC VOICE)

Features	Value
Sampling rate	16 KHz
Number of bits	16 bits
Number of Channels	1, Mono
Audio data file format	*.wav / Synthetic
# Speakers	19
# Language	3 (AR,FR and AN)
# speaker's gender	Male& female
# Data augmentation algorithms	2
# Total number of tokens Condition of noise	11560 normal life for direct recording corpus
Preemphased	$1 - 0.97z{-}1$
Window type Hamming	25.6 ms
Frames overlap	10 ms

1) Data Composition

• *Synthetic Data:* The dataset includes over 5,000 synthetic recordings created using a text-to-speech (TTS) system. To add variety and realism, different computer-generated voices were used, and techniques like adding noise, shifting the audio, and altering the timing were applied.
• *Normal Data:* We have over 6,000 recordings of real people speaking,

captured in everyday settings with a 16 kHz sampling rate and 16-bit quality. Each sentence was recorded once, reflecting natural variations in speech.

2) Data Collection Methodology

• *Normal Voice Recordings*: These recordings were made in various real-life settings, including quiet rooms, outdoor areas, and other everyday environments. This approach ensures a broad range of background sounds and conditions.

• *Synthetic Voice Recordings:* For synthetic data, we used a text-to-speech (TTS) platform to generate high quality speech with different synthetic voices. This provides a range of speech patterns for greater variety.

3) Data Augmentation Techniques

To further enhance the diversity and robustness of our corpus, several data augmentation techniques were applied to both synthetic and natural speech data. These techniques included the introduction of background noise and the use of signal processing methods such as rolling (shifting). Such augmentations were crucial in generating a more varied dataset that effectively simulates different real-world conditions and speaker variations.

1) Noise Addition: Incorporating background noise to replicate various environmental conditions.

2) Rolling: Adjusting the timing of the audio to vary the rhythm of the speech. The integration of both synthetic and real recordings results in a comprehensive dataset that is well-suited for training and evaluating language identification models, thereby improving their performance in practical applications.

B. Results and Discussion

In this section, we present and discuss the results of our study on spoken language identification using various features, including acoustic, temporal, and rhythmic elements. We assessed the performance of MLP, SVM, and LSTM models in three scenarios: identifying natural speech, synthetic speech, and a combination of both. Our main goal is to understand how synthetic speech affects identification accuracy compared to natural speech. The results offer insights into how different features and model types handle the differences between synthetic and natural speech.

1- Synthetic spoken language identification System

Based on the data presented in Table III, the following conclusions can be drawn:

• *MLP:* Achieves the highest overall performance, with a perfect score of 100 in both MFCC and Mel features, as well as a combined score of 100% across all features. This demonstrates the MLP model's superior effectiveness when utilizing these features, resulting in the best overall accuracy.

• *SVM:* Shows solid performance with a combined score of 94.01%, excelling particularly with contrast and chroma features. However, its performance with MFCC and Mel features is not as strong as that of the MLP and LSTM models.

• *LSTM:* Exhibits strong overall performance with a score of 99.93%, particularly excelling in processing MFCC and Mel features. However, it is less effective with chroma, contrast, and Tonnetz features compared to the other models.

TABLE 3. SYNTHETIC SPOKEN LANGUAGE IDENTIFICATION SYSTEM RESULTS

Model/Feat	MFCC	chroma	Mel	contrast	Tonnetz	ALL
MLP	**100**	62.92	**100**	66,23	44,44	**100**
SVM	85,89	65,98	83,58	80,4	47,13	94,01
LSTM	98.65	39.57	99.87	59.17	37.51	99.93

In summary, both MLP and LSTM models demonstrate superior effectiveness across various audio features, making them highly suitable for complex language recognition tasks. While the LSTM model achieves impressive overall accuracy, the MLP excels particularly with MFCC and Mel features. In contrast, the SVM model performs optimally with certain features like contrast and chroma but shows weaker overall results.

For tasks requiring strong performance across a variety of features, the LSTM or MLP models are the better choices. However, when focusing on specific feature types, the SVM model may be more effective. Furthermore, the influence of synthetic speech on spoken language identification systems is significant. The quality and characteristics of synthetic voices can vary substantially, affecting the accuracy of the system. High-quality synthetic

voices contribute to improved model performance by providing clear and consistent speech data, while lower-quality voices may introduce distortions, challenging the robustness of the models. Therefore, integrating high-quality synthetic voice data can enhance the system's ability to process diverse audio inputs.

2- Human-Recorded Voice-Based Spoken Language Identification System

The results shown in Table 4 outline the performance of several machine learning and deep learning models (MLP, SVM, and LSTM) using various feature sets, including MFCC (40), chroma (12), Mel (128), contrast (7), Tonnetz (6), and a combination of all features (ALL). The evaluation is centered on classification accuracy for language identification based on multilingual vocal commands.

TABLE 4. LANGUAGE IDENTIFICATION USING RECORDING SPEECH RESULTS

Model/Feat	MFCC	chroma	Mel	contrast	Tonnetz	ALL
MLP	48,66	37,37	77,41	37,1	42,99	57,71
SVM	43.95	34,18	52,3	46,95	46.95	44,65
LSTM	53,8	32,76	*74,35*	35,27	38	53,15

The results show that the Mel spectrogram (Mel(128)) consistently achieves the highest accuracy across all models, while the combination of all features (ALL) significantly improves performance for the MLP model.

- The MLP model stands out with exceptional performance, particularly when using the Mel spectrogram and the combined feature set. The LSTM model also performs well with the Mel spectrogram, though it is less effective with the combined feature set. In comparison, the SVM model shows lower performance across all feature sets.
- Variability in natural voice recordings impacts model performance, and an imbalanced dataset complicates the training process, potentially resulting in biased predictions.
- While data augmentation techniques can help address dataset imbalance, they do not fully capture the complexities of natural voice recordings. This study underscores the significance of feature selection, with the Mel spectrogram being the most effective. It also highlights the potential of MLP and LSTM models for spoken language identification tasks.

3- Impact of Synthetic Voices on Multilingual Language Identification

This study explores the effect of synthetic voice data on the performance of language identification systems. By comparing results from datasets consisting of synthetic, natural, and combined voices, we examine how the consistency and controlled nature of synthetic voices can significantly influence model accuracy when compared to natural speech recordings. This section discusses the advantages and challenges of using synthetic voices, offering insights into the complexities of language identification across different types of speech data. Based on the results presented in Table 5, the following observations can be made:

TABLE 5. IMPACT OF SYNTHETIC SPEECH ON SPOKEN LANGUAGE IDENTIFICATION SYSTEM PERFORMANCES

Model/Feat	MFCC	chroma	Mel	contrast	Tonnetz	ALL
MLP	71,15	44,73	*88,13*	49,14	41,5	76,17
SVM	56,6	51,56	64,12	51,41	44,47	64,18
LSTM	70,74	37,8	71,58	44,09	37,34	74,94

1) The combination of synthetic and natural voice data generally leads to improved performance across all models compared to using only natural voice data. However, it does not reach the performance levels achieved with synthetic data alone. This suggests that while synthetic data offers a consistent pattern that enhances model performance, the variability introduced by natural voice data continues to pose challenges.

2) The MLP and LSTM models particularly benefit from this combined approach, especially when using features like Mel and MFCC. This indicates that these models can effectively learn from both types of data. However, the performance improvement is not as pronounced as when using synthetic data exclusively, likely due to the additional variability of natural speech.

3) The SVM model also shows some improvement with the combined data, but its performance remains lower compared to the MLP and LSTM models. This suggests that SVMs may not be as effective in handling complex, variable data sets

This analysis underscores that while synthetic voice data can substantially improve language identification system performance, especially for MLP and LSTM models, the variability in natural speech introduces challenges that synthetic data alone cannot resolve. Combining synthetic and natural data provides a more balanced approach, but further optimization, such as advanced data augmentation or more refined feature engineering, may be needed to achieve the best performance across all conditions, particularly with short sentences.

4) Discussion*:* The results of this study highlight the significant role that different types of speech data—synthetic, natural, and combined—play in the performance of spoken language identification systems. The comparison across multiple models (MLP, SVM, and LSTM) reveals distinct patterns in how these models adapt to and perform with varying types of input data.

Performance of Different Models

In terms of overall performance, the MLP and LSTM models consistently outperform the SVM model across most feature sets. The MLP model, in particular, shows remarkable accuracy when utilizing Mel spectrogram and MFCC features, with an impressive performance boost observed when combining features like Mel spectrogram and the complete feature set (ALL). The LSTM model, while also strong, shows a slight decline in performance when dealing with combined feature sets, though it still maintains a high level of accuracy with Mel spectrograms. On the other hand, the SVM model, which excels in handling specific features such as contrast and chroma, underperforms relative to MLP and LSTM when faced with the complexity of multiple feature sets. This suggests that MLP and LSTM models are better equipped to handle the multidimensional nature of language identification tasks, especially when diverse speech features are involved. These findings align with existing literature, which indicates that deep learning models (such as MLP and LSTM) are particularly well-suited for complex tasks where data is highly variable and feature-rich.

Impact of Synthetic Voices

A key component of this study is the exploration of synthetic voice data in comparison with natural voice data. As synthetic voices exhibit uniformity and controlled characteristics, they contribute to consistent model performance. The results indicate that synthetic speech data leads to high accuracy,

particularly for MLP and LSTM models, which benefit from the predictability and clarity of synthetic voices. This finding is consistent with previous studies that have demonstrated the advantage of using synthetic voices to improve the reliability and consistency of speech recognition models. However, when combining synthetic and natural voice data, the performance improves compared to using only natural voice data, but it does not surpass the results obtained with synthetic data alone. This highlights an important point: while synthetic data provides an excellent foundation, natural speech recordings introduce variability—such as accents, speaking rates, and environmental noise—that synthetic data alone cannot capture. This variability poses challenges for the models, particularly for the MLP and LSTM models, whose performance gains from the combined data are less pronounced. This also reflects the inherent limitations of synthetic voices, which, despite their advantages, do not fully replicate the diversity and nuances found in natural speech.

Role of Data Augmentation and Feature Selection

Another notable finding is the role of data augmentation in addressing dataset imbalance. While data augmentation techniques can mitigate some of the issues associated with imbalanced datasets, they are not fully capable of capturing the complexity of natural voice recordings. This limitation suggests that synthetic voice data, though helpful, is not a perfect substitute for real-world speech data. The study emphasizes the importance of feature selection in improving model performance, with the Mel spectrogram emerging as the most effective feature for language identification across all models. This aligns with the growing body of research that highlights Mel spectrograms as a robust feature for speech recognition tasks. In addition, while combining synthetic and natural data provides a more balanced dataset, the challenge of optimizing models for performance in both controlled (synthetic) and variable (natural) environments remains. Advanced data augmentation techniques, along with further refinements in feature engineering, may be required to improve model performance, particularly in situations involving short sentences where the amount of available data may be limited.

5. CONCLUSION AND FUTURE DIRECTIONS

This research investigates the impact of synthetic speech on spoken language identification, evaluating the performance of various models—MLP, SVM, and LSTM—across natural, synthetic, and mixed speech types. The results demonstrate that Mel spectrograms yield the best outcomes, with both MLP and LSTM models achieving the highest accuracy levels. The highest accuracy, 100%, was achieved using Mel spectrograms, followed by MFCCs, which also reached 100% in synthetic speech. Interestingly, combining all feature sets did not always result in improved performance, highlighting the need for careful feature selection. The study further addresses the challenges posed by the variability in natural speech and imbalanced datasets, suggesting the need for robust data augmentation strategies. Looking ahead, future research could focus on expanding the dataset with more speakers and vocabulary, exploring advanced data augmentation techniques, tackling the complexities of dialects—which are more nuanced than language-level differences—and experimenting with deep learning models to boost system performance.

REFERENCES

Ambili, A.R. & Roy, R. C. (2023). The Effect of Synthetic Voice Data Augmentation on Spoken Language Identification on Indian Languages. *IEEE Access.*

Alashban, A.A. et al. (2022). Spoken language identification system using convolutional recurrent neural network. *Applied Sciences*, 12(18), 9181.

Alshutayri, A. & Albarhamtoshy, H. (2011). Arabic spoken language identification system (ASLIS): A proposed system to identifying modern standard Arabic (MSA) and Egyptian dialect. *IInformatics Engineering and Information Science: International Conference, ICIEIS 2011, Kuala Lumpur, Malaysia, November 14-16, 2011. Proceedings, Part II*, pp.375-385. Springer Berlin Heidelberg.

Biswas, M., et al. (2023). Automatic spoken language identification using MFCC based time series features. *Multimedia Tools and Applications*, 82(7), 9565-9595.

Duffy, S.A. & Pisoni, D.B. (1992). Comprehension of synthetic speech produced by rule: review and theoretical interpretation. *Language and Speech*, 35(4), 351-389.

Ganapathy, S. et al. (2014). Robust language identification using convolutional neural network features, *Interspeech*, pp. 1846-1850.

Gelly, G. & Gauvain, J.L. (2017). Spoken Language Identification Using LSTM-Based Angular Proximity. *Interspeech*, pp. 2566-2570.

Jothilakshmi, S., Ramalingam, V. & Palanivel, S. (2012). A hierarchical language identification system for Indian languages. *Digital Signal Processing*, 22(3), 544-553.

Kumar, P. et al. (2010). Spoken language identification using hybrid feature extraction methods. *arXiv preprint arXiv*:1003.5623.

Kumar, S. S. & Ramasubramanian, V. (2005). Automatic language identification using ergodic-HMM. In Proceedings. (ICASSP'05), *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005*. (Vol. 1, pp. I-609).

Maguolo, G. et al. (2021). Audiogmenter: a MATLAB toolbox for audio data augmentation. *Applied Computing and Informatics.*

Manchala, S. et al. (2014). GMM based language identification system using robust features. *International Journal of Speech Technology*, 17, 99-105.

Mc Fee, B. et al. (2015). librosa: Audio and music signal analysis in python. In *SciPy*, pp. 18-24.

Sarmah, K. & Bhattacharjee, U. (2014). GMM based Language Identification using MFCC and SDC Features. *International Journal of Computer Applications*, 85(5).

Sefara, T.J. et al. (2019). HMM-based speech synthesis system incorporated with language identification for low-resourced languages. *In 2019 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD)*, pp. 1-6.

Singh, G. et al. (2021). Spoken language identification using deep learning. *Computational Intelligence and Neuroscience*, 2021(1), 5123671.

Wazir, A.S.B., et al. (2020). Spectrogram based classification of spoken foul language using deep CNN. *In 2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*, pp. 1-6.

Wicaksana, V.S. & Kom, A.Z.S. (2021). Spoken language identification on local language using MFCC, random forest, KNN, and GMM. *International Journal of Advanced Computer Science and Applications*, 12(5).

Zazo, R. et al. (2016). Language identification in short utterances using long short-term memory (LSTM) recurrent neural networks. *PloS one*, 11(1), e0146917.