ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

Expert Systems with Applications 00 (2023) 1–20

Speech Recognition

Acoustic & Language Agnostic Gender Detection

Or Haim Anidjar[a,b,c,d,*], Itamar Casspi[a], Sivan Cohen[a], Moriah David[a], Firas Naamneh[a], Hodaya Turgeman[a], Amit Waizman Israel[a], Daniel Zaken[a], Stav Zilber[a], Roi Yozevitch[a]

[a]*School of Computer Science, Ariel University, Golan Heights 1, 4077625, Ariel, Israel.*
[b]*Ariel Cyber Innovation Center, Ariel University, Golan Heights 1, 4077625, Ariel, Israel.*
[c]*Data Science and Artificial Intelligence Research Center, Ariel University, Golan Heights 1, 4077625, Ariel, Israel.*
[d]*Kinematics and Computational Geometry Lab (K&CG), Ariel University, Golan Heights 1, 4077625, Ariel, Israel.*

Abstract

This study proposes an agnostic method for identifying the gender of the speaker from an audio clip in a noisy environment. We perform two different processes on audio clips as a Mel-Spectrogram and once with the Wav2Vec2 acoustic model emission and examine the advantages and disadvantages of each method. We present a series of experiments across five different languages English, Arabic, Spanish, French, and Russian containing male and female audio clips. An analysis process of the languages is carried out while examining their agnostic characteristics against the background of a model of five languages, The goal of our study is to distinguish the gender of the speaker based on an audio clip regardless of language or complex background noise such as nightclubs or stadiums. The experimental results indicate that the performance evaluation of the traditional spectrogram method achieved better results compared to the Wav2Vec transformer method. For the Russian language, the spectrogram method achieved an accuracy of 99%, While the wav2vec transformer method achieved only 89% accuracy. Another experiment was tested in several different types of environments: noisy, silent , noisy and silent environment. The experimental results show that a model trained in a noisy and silent environment exhibited better accuracy compared to the others. In addition,the results of the experiment indicate that a model trained on data from a wide variety of languages yielded higher results. The research findings highlight important insights for developing a more reliable and accurate system.

Keywords: Wav2Vec 2.0, Mel-Spectrogram,Language Agnostic ,Gender identification.

1. Introduction

It is common to say that in human communication most of the information is not verbal DeVito et al. (2000). Beyond the words that someone says, there are also many details about the person who speaks, like body language, intonation, and more. Humans vary pitch, speed, loudness, tone, rhythm, body motions, and facial expressions. It all helps convey additional information and provide a sense of the message DeVito et al. (2000); Key (2011).

When it comes to Automatic Speech Recognition (ASR) systems, the understanding that human communication has **more** than just words is crucial. On ASR, the audio file is converted to speech, i.e., meaningful words, referred to as transcripts. However, there is an additional layer beyond the transcription that is equally important - metadata; it

*Corresponding author: Or Haim Anidjar, School of Computer Science, Ariel University, Golan Heights 1, 4077625, Ariel, Israel.

Email addresses: `orhaim@ariel.ac.il` (Or Haim Anidjar), `Itamarcasspi@gmail.com` (Itamar Casspi), `Sivancohen0987@gmail.com` (Sivan Cohen), `Moriah.David24@gmail.com` (Moriah David), `firas457@gmail.com` (Firas Naamneh), `hodaya.tor@gmail.com` (Hodaya Turgeman), `amitw6363@gmail.com` (Amit Waizman Israel), `danielzaken5@gmail.com` (Daniel Zaken), `stavzilber@gmail.com` (Stav Zilber), `roiyo@ariel.ac.il` (Roi Yozevitch)

contains various elements such as speaker identity, speech intonation, emotions, and more. By extracting this metadata from RAW audio, ASR systems will have a greater understanding of the speech and the conversation will be easier to comprehend.

The human voice contains personality traits such as: Language, age, the unique shape of the vocal system Pahwa & Aggarwal (2016), accent, rhythm, and genderUddin et al. (2022) (for a detailed survey, see Jahangir et al. (2021)). In this paper, we focus on gender classification. Gender classification by voice refers to the process of determining gender based on the analysis of various parameters extracted from a vocal sample and classifying it as a female or male voice.Submitter et al. (2021).

There are many motivations for a gender classification system Nur et al. (2022). The first example focuses on the privacy of women in clubs. Privacy and safety are important to people in public spaces, for example in clubs where there are many cases of privacy violations and harassment of people, especially women Mellgren et al. (2018). Thus, creating an environment for women's well-being, comfort, and security is crucial, especially in clubs, where the music and loud noise make it difficult to detect that something wrong is happening Cannon et al. (1998). Implementing a system that classifies the person's gender at the entrance to the women's toilets and allows entry only to women is important in this context. By having a reliable classification system, women will be able to feel confident that the spaces intended for them will remain only for them. A significant advantage of a voice recognition system in public places is creating a balance between privacy and security by providing an effective identification tool that avoids the need for visual surveillance that causes privacy violations (e.g., security cameras).

Discerning the gender of a speaker solely based on the language used can be a deceptive task due to various linguistic nuances. These challenges are further compounded when the speakers originate from different regions or speak different dialects, even within the same language. For instance, with its diverse dialects, Arabic presents varied pronunciations and sentence structures that could hinder effective communication Farghaly & Shaalan (2010).

In essence, each language possesses distinct characteristics and components that render it unique. Some languages are recognized as "gender-neutral," meaning that the language avoids making presumptions about the social gender or biological sex of the people being referred to in speech Wright (2002). English, for instance, is a gender-neutral language where words, phrases, and sentences can be utilized without specifying the gender of the person or group being referred to Hord (2016).

However, some languages are not gender-neutral and exhibit sexual differentiation in words. The structure of these languages often indicates gender - for example, adding a certain suffix to a word might specify if it belongs to a male or a female. Hebrew and Arabic are such languages, with masculine and feminine forms for names, nouns, adjectives, and pronounsCowley & Kautzsch (1910).

Despite these linguistic characteristics, the use of language itself poses significant challenges for reliable gender detection. In gender-specific languages, individuals may choose to use words and sentence structures traditionally associated with the opposite gender, either due to personal preference or an intent to mislead, thus creating potential for false gender identification. Furthermore, in gender-neutral languages, the absence of gender-specific language rules does not facilitate the identification of the speaker's gender.

Therefore, due to the linguistic variations in gender-neutral and non-neutral languages and the potential for misleading gender identification, the words themselves are not a reliable source for gender classification. This necessitates the development of a system that can effectively detect gender based on acoustic properties rather than the content of the speech.

Therefore, to face these challenges, we must build a robust and reliable system. We can do this using Acoustic and Language Agnostic Gender Detection (ALAGD). In acoustic gender recognition Simonović et al. (2021), we determine the speaker's gender according to more basic characteristics such as pitch, voice quality, and speed. Language-agnostic gender recognition allows us to accommodate methods for owners of different languages without relying on linguistic and cultural characteristics Pikuliak et al. (2021).

For example, we can look at the feature that men tend to have a lower voice than women. To build a language-agnostic gender recognition system Anidjar et al. (2023), we need to treat features in a universal way and identify more basic and general acoustic features.

There is great importance in collecting the data and adapting it to the research goals. A proper data processing process will help us map and extract the relevant features from an audio clip as part of developing a reliable and robust ALAGD model. To do this, we first tried speech recognition using the wav2vec method.Baevski et al. (2020) wav2vec is a deep learning model Sun et al. (2023). Based on the Self Supervised Learning (SSL) method of speech representations characterized by two central processes,In the first stage, preliminary training is performed during which unlabeled speech data are used to learn to predict speech signal segments and capture important features found there.

The Choi et al. (2022) demonstrates the use of SSL method In the second step, we tune the model using a smaller amount of data specifically labeled according to the target. Through this step, we can improve the performance of the model. There are other works that have used this method and achieved good performance Deschamps-Berger et al. (2022).

One of the central things that the Wav2vec 2.0 method provides in acoustic and language-agnostic gender recognition is that through the process of training the model in a wide variety of languages, the possibility and ability to extract features from a speech signal is given regardless of the language. Carrying out this process creates a much more robust and general approach that achieves a significant advantage.

Another method we will present in this article is the mel-spectrogram.Zolnay et al. (2005): This is a very popular method that performs analysis on audio clips using signal processing.

An audio segment consists of several sound waves at a single frequency. The process by which we convert the signal from the time domain to the frequency range is called a spectrogram. A special weights function is activated, in the middle of which you can extract important information about various properties of the sound waves, such as sounds, pauses between sounds and more. This is important information that we will use in the research.

In addition, the Mel-spectrogram is used as an input to the CNN network Gupta et al. (2022). The CNN learns the main features from the Mel-spectrogram to identify patterns and features in the audio. The CNN includes convolutional layers that mix multiple filters on the mel-spectrogram. In this way, we will be able to perform feature extraction in a precise and correct manner, which is a necessary step in building the model.

This study aimed to develop an agnostic language model to improve gender detection from the voice segment of a speaker regardless of the speaker's language. The process begins with the creation of a dataset from Common Voice. The data is audio segments of male and female speakers in English, Spanish, French, Russian, and Arabic that are three seconds long. The first experiment was divided into two parts and the training and the test was performed on each language separately. It performed once with the audio segments as a mel spectrogram and once with the emission of the Wav2Vec2 acoustic model. The aim of using these two methods is to understand which one has better performance. In the second experiment, the data were divided into groups so that each group has a combination of three different languages and was performed only with the emission of the Wav2Vec2. For each group, the model was trained with the three languages and the test was performed on the remaining two languages. In addition, this experiment was also performed on the five languages together with mel spectrogram and Wav2Vec2. The third experiment was similar to the second, but this time it was meant to check the model's durability while audio segments combined with trance music. The experiment was divided into two parts and was performed with mel spectrogram and Wav2Vec2. The first was trained and tested with 80% original data and 20% from the rest data with trance music pieces. The second was trained with 100% original data and the test was performed with 100% of data with trance music pieces.

1.1. Our Contributions

The contributions of this study are summarized as follows: This study makes several distinct contributions to the understanding and application of deep learning in Natural Language Processing (NLP). An insightful comparison of Transformer and Spectrogram data techniques when used in conjunction with Convolutional Neural Networks (CNN) is provided, exploring their individual strengths and weaknesses across a diverse set of languages. Language agnosticism within multilingual systems is also investigated, examining the dynamics and interplay between various languages and various models that were trained on different languages and different numbers of languages. Finally, these techniques are applied in a practical scenario, studying the impact of different noisy environments on model performance. (i.e. nightclub, stadium etc.)

- **Special use case** - An investigation was conducted into the effectiveness of generic models trained in noisy environments versus those trained in silent settings. This comparison was executed by testing the models on both silent and noisy datasets, thus providing direct empirical evidence of the advantages and drawbacks associated with training a model for a specific scenario, such as a noisy environment.

- **Wav2Vec versus Spectrogram** - A comparative analysis was made of two distinct data types: wav2vec's emission and a conventional spectrogram method. Upon comparing the two, it was discovered that the spectrogram method yielded superior results relative to wav2vec. However, it was also noted that the runtime performance of wav2vec significantly surpassed that of the spectrogram method, manifesting an edge both during the preprocessing stage and throughout the training stage.

- **Language Agnosticism** - Another major aspect of this study was the examination of language agnosticism and how multilingual models interact with various languages. By exploring different combinations of languages, insights were gained about the dynamics within multilingual systems and the impact of individual languages on overall model performance.
- **Cost-Effectiveness and Resources** - The cost-effectiveness and suitability of each method for different use cases were considered. It was noted that while the wav2vec method offers advantages in terms of speed, it may be less accurate. On the other hand, the traditional spectrogram method may have longer runtime, but it tends to be more accurate.
- **Comparing models accuracy on different languages** - The study also explored how different languages can influence the model's accuracy in gender prediction tasks. It was discovered that the performance varies depending on the language used, shedding light on how linguistic features can affect prediction tasks.

1.2. Paper Structure

The remainder of this paper is structured as follows:

- Section 2 A review of studies that contributed to the project and offered different techniques that help identify a speaker by gender and different tools that contributed to the effectiveness of the model. Overview of Wav2Vec2 architecture and different models from deep learning.
- Section 3 Discusses the setup used to develop the forecasting model. This paper uses the Common Voice set of work including five different languages English, Arabic, Spanish , French and Russian.
- Section 4 Discuss in depth the approach proposed in this article, the actions we performed and the models we used exploitation of the Wav2Vec2 architecture on the audio data and combining languages in order to create a robust and reliable model.
- Section 5 presents the results of the experiment consisting of an evaluation of the predicted model using Metrics that provide useful information about the performance of the model and its level of functioning in Section 5.2, and the presentation of the three experiments carried out in the study and the results of the comparison between our approach and different baselines in Section 5.3;
- Finally, Section 6 concludes and summarizes this paper.

For ease of reading, we provide a list of abbreviations in Table 1.

2. Related Work

Studies that contributed to the project through interesting techniques and tools Baevski et al. (2020): presents the Wav2vec 2.0 method that performs self-supervised learning from raw audio data for speech recognition. This method encodes speech audio using a multi-layer neural network. The study offers solutions and improvements that include a two-step pre-training process in which the model is trained on unlabeled audio signal data that allows the model to learn audio features relevant to speech independently. and targeting the model using a smaller amount of labeled data that allows targeting the model to a more specific goal. In addition, the method uses deep neural layers to extract more complex information about the features from the speech signal segments.

One of the main advantages of this method is the use of a model that was trained independently from the data of the audio file without the need for certain data or manual analysis in advance. But a notable disadvantage is the fact that this method requires extensive computational resources and a lot of time to bring about advanced and improved

Abbreviation	**Meaning**
ALAGD	Acoustic and Language Agnostic Gender Detection
ASR	Automatic Speech Recognition
CNN	Convolutional Neural Network
FFT	Fast Fourier Transform
LDA	Linear Discriminant Analysis
MACE	Mean Absolute Class Error
MAE	Mean Absolute Error
MFCC	Mel Frequency Cepstral Coefficient
MLNN	Multi-Layer Neural Network
NN	Neural Network
SSL	Self Supervised Learning

Table 1. List of Abbreviations. Sort by alphabetic order

results. The paper Zolnay et al. (2005) focuses on improving the performance of automatic speech recognition systems by combining different acoustic features.

The paper examined two main methods for combining the acoustic characteristics: LDA based combination Haeb-Umbach & Ney (1992) and log-linear based combination Beyerlein (1997).

The paper suggests a combination of different voice characteristics in order to improve the accuracy of identification. They use a combination of two key features, which are: Spectrogram (Mel-Spectrogram) this characteristic refers to the range of frequencies and times in sound. It shows the different frequency spectrum within the voice and the changes in time. Mel-frequency cepstral coefficients (MFCC) this characteristic refers to the centrality of the frequencies in the voice. It presents the frequency battles in voice effectively. Özcan & Kayıkçıoğlu (2021) From the results of the study it emerged that a combination of MFCC and Mel Spectrogram features. resulted in improved identification of the system. In addition, research shows that the log-linear combination of features also strengthens the system's ability to identify and classify. One of the most prominent advantages is a combination of different sound characteristics that improves the Ability to identify the automatic voice systems in the areas affected by acoustic problems, but the combination of the voice characteristics and possibilities to carry information and a large data size that may increase resource consumption

Other interesting studies dealing with the gender identification of the speaker. the paper Chachadi & Nirmala (2022) describes the use of a Neural Network (NN) Wu & Tsai (2011) for gender recognition based on human voices. The main goal was to develop a model for gender recognition from an audio segment using a neural network and to examine how by adding different feature extraction techniques such as MFCC and mel-spectrogram the recognition can be improved. The results showed that the combination of MFCC and mel features achieved the highest accuracy of 94.32% on the training set. The research results highlight the ability of the NN model to effectively capture gender-related features from speech signals. However, there are potential limitations to the model, particularly its reliance on acoustic features. Acoustic features alone may not fully capture these complex gender-related aspects. For example other factors such as regional accents, cultural backgrounds, or personal traits. In our research we use the CNN method and not the NN. CNN learns to automatically extract patterns and hierarchical features from input data it has the ability to efficiently identify spatial features, making it particularly useful for analyzing audio signals.

In Livieris et al. (2019), in order to identify the speaker's gender, a semi-supervised algorithm Karlos et al. (2019), called iCST-Voting for the gender recognition by voice. This algorithm is a combination of all the most useful self-labeling algorithms: self-training, co-training and tri-training. Meel & Vishwakarma (2021), Grolman et al. (2022). two experiments took place in two distinct phases: In the first phase, the performance of iCST-Voting was evaluated and compared against its individual component self-labeled algorithms, namely Self-training, Co-training, and Tri-training. Additionally, the performance of iCST-Voting was also compared against the state-of-the-art self-labeled algorithms including SETRED, Co-Bagging, Democratic-Co learning, and Co-Fores. While in the second phase, the goal was to compare the performance of the proposed algorithm iCST-Voting with classical supervised algorithms. The article proposes an approach that combines these different types of classifiers in order to achieve more accurate classifications than classical supervised algorithms. The article presents a set of experiments and shows that the iCST-Voting algorithm achieves higher results than other machine learning algorithms and reaches an accuracy of 98.23%.

Another article Perry et al. (2001), focused on analyzing the specific acoustic characteristics Xie & Zhu (2019) present in the speech and voices of children. The objective was to determine how these characteristics influence the ability of listeners to correctly identify the gender of the children. In the first experiment, researchers collected vocal recordings and physical measurements from children belonging to different age groups. Each group included an equal number of boys and girls. The speech samples used in the study included seven vowels from the American English language. The researchers measured various aspects of the speech samples, including the fundamental frequency (f0) and formant frequencies (F1, F2, F3) present in these vowel syllables. In the second experiment, a group of 20 adults were involved. They listened to the recorded syllables produced by the children in the first experiment. The adults were then asked to rate the gender of the speakers using a six-point gender rating scale. The results of the experiments revealed that These acoustic features become more apparent as children grow older. That is, the older the speaker, the more frequencies that contribute to the distinction between the sexes are created. This implies that changes in physical size can influence vocal characteristics. The disadvantage of this study is that its data set is of children only, in the current study the audio clips are of adults.

In paper Alkhammash et al. (2022),introduced a gender voice recognition model that utilizes a stacked ensemble Agarwal & Chowdary (2020) approach. The research was done by using four machine learning algorithms that help to find the correct classification, k-nearest neighbor (KNN) Gou et al. (2022), support vector machine (SVM) Alcaraz et al. (2022), stochastic gradient descent (SGD), and logistic regression (LR) as base classifiers and linear discriminant analysis (LDA) Moscatelli et al. (2020) as meta classifier. The main objective of this article is to showcase the implementation of a stacked ensemble model in gender voice recognition. The performance of the proposed model was compared to traditional machine learning models, and it achieved the highest accuracy rate of 99.64%. In this stacked model, five machine learning models were utilized. Four of these models were used as the base classifiers, while one model served as the meta classifier. To obtain the most accurate predictions, data preprocessing techniques and k-fold cross-validation were employed. The disadvantage of this article is that stacked ensemble models heavily rely on having diverse and representative datasets, however the amount of data for the given article is not extensive. In the current study, the data set is wide and consists of different languages in order to produce a variety of audio samples.

Now we will expand on some language agnostic gender detection projects.

In Janeva et al. (2022) they explore different approaches to voice recognition using machine learning and deep learning models. They focus on predicting gender, age range, and combined gender and age range using a multilingual dataset. They use five machine learning models and a CNN deep learning model Gu et al. (2018) for training and evaluation. The dataset contains audio recordings from various languages, including English, Italian, French, Spanish, Russian, Portuguese, and others. The aim is to understand the similarities and differences in voice characteristics across different languages. The dataset is imbalanced, with a larger number of male speakers compared to female speakers. The results show that Random Forest Prinzie & Van den Poel (2008) is better than other models in accuracy of more than 90% for all three classification tasks. They use SHAP method Lundberg & Lee (2017); Lundberg et al. (2018) which is based on the concepts of game theory, it has been used for multiclass classifiers to increase the reliability of the prediction results. The results highlight the influence of specific features on gender and age range prediction. Although this, Wav2Vec2 has advantages over the Random Forest model when working with sequential data like audio segments. It can directly process raw waveforms, and handle variable length inputs, making it a more suitable choice for audio segment tasks.

The paper Lastow et al. (2022) focuses on developing machine learning models that predict the gender of a speaker and his age using their voice samples, particularly in multilingual settings, to improve conversational interactions.

They created four different datasets with data extracted from the Common Voice project to compare monolingual and multilingual performances. On gender classification, they reached a macro average F1 score of 96% in both a monolingual and multilingual setting. For age classification, using classes with a size of 10 years, they reach a macro average mean absolute class error (MACE) of 0.68 on monolingual datasets and 0.86 on multilingual datasets. With the WavLM model Zhao & Zhang (2022), and with English TIMIT dataset Zoughi et al. (2020), they reached a mean absolute error (MAE) of 4.11 years for males and 4.44 for females in age estimation, and their fine-tuned UniSpeech-SAT model Wang et al. (2021) achieves an amazing accuracy of 99.8% for gender classification. By use of multilingual datasets, the model can learn and recognize independent features that are relevant to age and gender classification in an efficient way. The Wav2Vec2 model has advantages over the WavLM and UniSpeech-SAT models in several aspects. It can process raw waveforms, and handle variable length inputs. In addition, WavLM and UniSpeech-SAT models primarily focus on language modeling tasks using text inputs and may not be as optimized for gender detection tasks.

In Kalaycı & Doğan (2020), focused on determining the gender of speakers based on the acoustic characteristics of their speeches. The researchers compiled speech samples from seven different languages and analyzed the acoustic features of the speakers' voices to identify their gender. This paper is also language agnostic as its dataset contains seven different languages. In this article, both data mining and deep learning techniques were used. The article presents the use of different machine learning algorithms Dogan & Birant (2021) - Logistic regression, SVM, KNN, XGBoost Mushava & Murray (2022) and AdaBoost Huang et al. (2022). Among the methods, the highest accuracy was achieved using gradient boosting and random forest for classification. In addition to this, a deep learning method called Multilayer Perceptron (MLP) Arias del Campo et al. (2021) was used, with the help of which they reached a correct classification rate of 96.22% in their analysis. There is a similarity in this study to the current study since a variety of languages were used here as well. The advantage of the current study is that there is a variety of experiments in which each time the languages in the train and the languages in the test are changed, different experiments were conducted with more advanced techniques and each time a different number of languages.

Overall, The benefit of the present study is the use of five different languages containing lots of audio samples. At the beginning of the research, wav2vec was used, which is known for its speed compared to other techniques. The article describes five experiments of individual languages with emission, five experiments of individual languages with a spectrogram, four experiments of three emission languages, one experiment of all five languages and two experiments related to music in the background (one all music and one part music). Another thing, this study describes a language-agnostic system, regardless of the speaker's language, the model will be able to identify the speaker's gender.

3. Dataset

The study recognizes the importance of developing a ALAGD model that is trained in several languages,This can help the identification to be more accurate when any language is given and not necessarily a specific language. Thus, the Common Voice dataset was utilized. `https://commonvoice.mozilla.org/he` This dataset has been used in many studies on gender classification in recent years such as: Tursunov et al. (2021a), Alnuaim et al. (2022).

Common Voice is a project started by "Mozilla" with the aim of creating a free database for speech recognition software. People upload recordings of themselves speaking in all kinds of languages and also review recordings of other participants. Currently, Common Voice has 108 languages and 27, 142 recordings hours of people . Many of the recorded in the dataset also includes demographic metadata like age, sex, and accent. The current model is processed on 5 languages: English, Spanish, French, Russian and Arabic, among the 108 languages that the data contains.

Here is a graph that shows the percentages of recordings by men and women in the 5 languages that the study deals with. The "gender" label is divided into women, men and unknown.

Languages	**Male clips**	**Female clips**	**Percentage of male clips**	**Percentage of female clips**	**Average duration length**
English	787,886	270,199	46.63%	15.99%	5.097
Spanish	190,986	83,950	53.54%	23.53%	5.087
French	410,796	71,068	60.71%	10.50%	4.944
Russian	90,285	24,045	60.74%	16.18%	5.112
Arabic	17,807	14,836	23.22%	19.35%	4.082

Table 2. The table presents the distribution of male and female voices in the Common Voice dataset for various languages. It highlights the potential bias in representation between men and women.

As can be seen from the table ,there is a fairly big difference between the number of recordings in English, French, Spanish and Arabic.Also,in all languages there is a higher number of recordings of men than women, This problem is called ”gender bias” and it is a common problem. The current article solves the problem by reducing the number of recordings in languages with the highest number of recordings, and in addition, it compares the number of recordings between men and women.

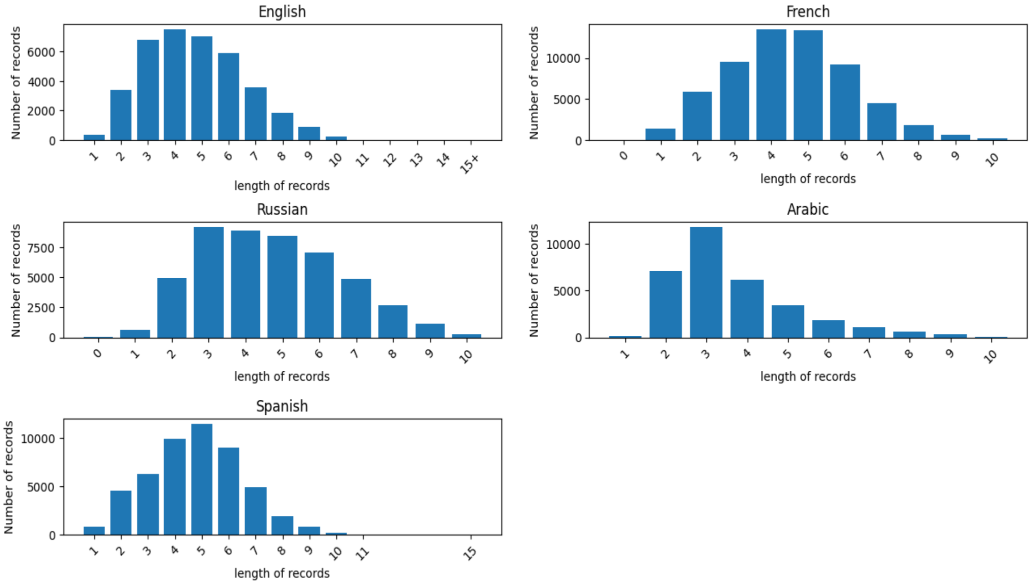

Figure 1. The figure presents the original length of the recordings from each language taken. The x-axis is the length of the recording and the y-axis is the amount of recordings of the same length.

In order to adapt the data more to the reality that the research is dealing with, the data consists recordings of 3 seconds from the original recordings.

3.1. Gender Bias

Gender bias in voice recognition systems is evident through the presence of biases favoring male voices in gender classification tasks. These biases arise due to imbalances in training data, where male voices are overrepresented compared to female voices. This leads to lower accuracy and higher error rates for female voices during gender classification. The underrepresentation of female voices in training data can result in a lack of robustness and fairness in recognizing and transcribing female speech, exacerbating gender disparities in speech technologies. Addressing this bias requires diverse and balanced training data, along with algorithmic adjustments and evaluation metrics that consider gender fairness. Shahbazi et al. (2023) In our research we face gender bias while using Common Voice dataset. To address the issue of gender bias, the Common Voice project takes steps to encourage equal participation from both male and female contributors. The dataset strives to have a balanced representation of voices across genders. However, the actual distribution of male and female voices can vary depending on the contributions received.

Languages	**male clips**	**female clips**
English	20,000	20,000
Spanish	30,000	30,000
French	30,000	30,000
Russian	24,000	24,000
Arabic	17,000	14,000

In our paper, we address this issue by ensuring an equal distribution of male and female audio clips, as demonstrated in the dataset section. We recognize the importance of mitigating gender bias and took steps to balance the representation of male and female voices in our research. By utilizing an equal number of audio clips from both genders, we aimed to minimize any potential biases that may arise due to imbalances in the dataset. This approach allowed us to conduct a more comprehensive and fair analysis, promoting equality and inclusivity in our study.

this article Garnerin et al. (2019) analyzes the gender representation in four central bodies of the French broadcasting, they experience the impact of the imbalance of the genders in television broadcasting and radio system on the ASR performance and discovered that women are underrepresented in terms of speaking. another article Walker et al. (2022) treated this problem as an advantage. They used speaker recognition systems suffering from the problem of gender bias and used to achieve a new attack. They want to disrupt orders by attacking while exploiting today's speech systems still suffer from gender biases. Gender bias is a significant aspect of representation bias, which is a broader concept encompassing various biases in data. In addition to gender bias, our paper also addresses language bias, highlighting the underrepresentation of certain languages such as Arabic. By acknowledging and tackling these biases, we aim to promote inclusivity, fairness, and accuracy in ALAGD technologies across different genders and languages.

4. Framework

In the following sections, the framework and methodology that was employed in the research is presented. The goal of this article was to investigate the effectiveness of different training approaches and preprocessing methods in gender classification models based on speech recognition. The dataset preparation in Section 4.1.1 involves different allocations of the audio segments to create different performances of similar models. In the preprocessing stage at Section 4.1.2 different methods were used, which are mel-spectrograms and Wav2Vec2.0 emissions where both are used for extraction of meaningful features. Mel-spectrograms and Wav2Vec2.0 emissions also give the model the ability to use image-processing techniques on an audio wave segment.

Next insight is provided about the model's architecture, which makes use of a CNN model, explained in further details in Section 4.2. Lastly in Section 4.3 the model verification testing and inference is described. Overall the following framework section sets the stage for deeper analysis of the experimental results in section 5.

4.1. Preliminary operations.

The following section discusses the preliminary operations of all three experiments, including dataset preperation 4.1.1 and the preprocessing 4.1.2 stage of the model.

4.1.1. Dataset preparation

The dataset used for all conducted experiments is the Common Voice Ardila et al. (2019) public dataset by Mozilla. In the preparation of the dataset for the three experiments ahead, gender bias was revealed in the amount of labeled data found in each class. Gender bias is a term highly spoken about in the speech recognition community, that relates to the fact that most datasets related to speech are highly unbalanced in terms of the amount of samples from men and from women. This imbalance can impact the capabilities of the model regarding recognition of speech segments where the speaker is of the female gender. To balance the dataset equal amount of data segments were used from each gender, at the cost of losing data quantity.

This paper discusses three experiments each conducted with a different goal. The first experiment tests the effectiveness of the Wav2Vec2.0 Baevski et al. (2020) emission as opposed to using a mel-spectrogram Tursunov et al. (2021b) of that same data segment. In the second experiment we have trained the model agnostically, with multiple languages in the training phase to create a more reliable model that will have competitive performance when getting input from a variety of language speakers, for example a place that is highly visited by tourists. The idea behind this method came from the fact that tones and speaking rate differ from one language to another, making it hard for even a human ear to identify the gender of the speaker when his language is unrelated to his. In the third experiment the model was trained with augmented data, where some data segments were overlapped with trance music. The reason for this augmentation is creating a more robust model that is less sensitive to noises, increasing its performance in noisier scenarios, such as restaurants, clubs and basically every public space that is naturally noisier. The choice of using electronic trance music is that it is non vocal, to avoid overlapping speech segments between the actual speaker and the singer. The motivation of this augmented and agnostic dataset is to build a robust and practical model that is rich in real life use cases.

4.1.2. Pre processing

In this research the usage of mel-spectrograms and Wav2Vec2.0 emissions are present. These are methods of converting audio signals into 2D matrix representations. Mel-spectrogram transforms the audio waves from the time-domain to the frequency-time domain, which allows us to see the magnitude of the mel-frequency component at a particular time frame. such conversion is depicted in figure 2. The Wav2Vec2.0 model converts audio waves into high-level vectorial representation of the audio segment, encoding various acoustic features, where vectors with similar features will be short in distance.

These conversions allow the usage of image processing techniques and to leverage the power of a CNN model for feature extraction.

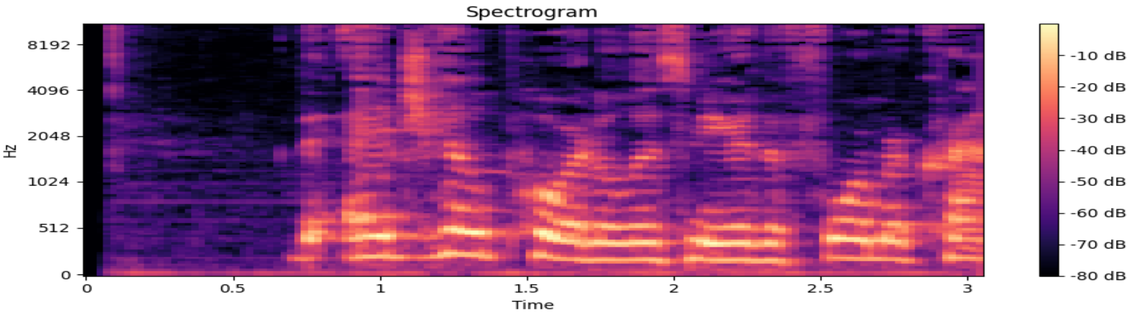

Figure 2. An mel-spectrogram example of a 3 second audio record from an English speaking woman.

- **Mel-Spectrograms** - Every audio segment from the data set was represented by a mel-spectrogram which is an image representation of audio signals which allows the usage of image processing methods while working with audio. The audio signals are divided into overlapping frames then applied a Fast Fourier Transform (FFT) to obtain the frequency domain representation and then a mel-scaled filterbank is applied to the power spectrum received by the FFT. A filter bank is a collection of filters designed to capture the energy distribution across different frequency bands in a way that better aligns with human perception of sound, and convolving the frequency band with each filter outputs the intensity of each band. Logarithmic scaling is then performed on the filterbank output energies to compress the range of values to emphasize lower energy while reducing high energy. The logarithmic scaling helps represent the perception of human hearing as it operates on a logarithmic scale rather than linear, and compresses the values of the energies to values that are more natural to the human ear. The mel-spectrogram is a 2D matrix where the columns represent the time frame of the original audio signals and the rows represent the mel frequency. Each element in the matrix represents the energy at a specific mel value frequency at a certain time window.
- **Wav2Vec2.0** - The Wav2Vec2.0 model was developed by the Facebook AI Research team, is a self supervised Hendrycks et al. (2019) framework for learning representations of raw audio signal waveforms. It is able to learn meaningful information from unlabeled data in a self-supervised training. The model was built in two stages, In the first stage the model learns how to predict masked portions of the speech segments. In the second stage, a contrastive loss function Wang & Liu (2021) was used to learn the discriminative representations of a segment and push them closer together, such that segments with similar contextual meaning will have similar embedding representation. The output extracted from the model captures unique characteristics of audio segments that have many different applications in the speech recognition world (such as intonations, temporal patterns and frequency).

4.2. Model architecture

The following description is of the model architecture used in this research. The model consists of six convolutional layers with varying number of kernel sizes and filters to perform feature extraction and classification to identify the gender of the speaker based on audio data. The configuration of each layer is designed to extract meaningful and accurate representation of the speaker. The ability of the model to learn hierarchical representation of the speaker that comes from the layering of the convolutional layer, contributes to the power of the model to accurately predict the gender of the speaker.

1. The first convolutional layer consists of 96 filters with kernel and stride size of 7x7 and two respectively. Following the first layer a batch normalization and max pooling layer with kernel and stride size of 3x3 and 2.
2. The second convolutional layer consits of 256 filters with a kernel and stride size of 5x5 and two respectively. Following this layer are similar batch normalization and max pooling layers as the first.
3. The third convolutional layer consists of 384 filters with kernel size of 3x3 and stride of 1. Following that is a batch normalization.
4. The fourth convolutional layer consists of 256 filters with kernel size of 3x3 and stride of 1. It is also followed by batch normalization.
5. The fifth convolutional layer is the same as the previous, but followed by a batch normalization and a max pooling layer with kernel and stride size of 5x3 and 3 respectively.
6. The sixth and final convolutional layer consists of 4096 filters with kernel size of 9x1 with no padding, followed by dropout to prevent overfitting.

Following the convolutional layers, the model uses adaptive average pooling to aggregate spatial values to a more compact representation. After that comes two dense layers, the first with 1024 units and the second dense layer predicts the gender of the speaker.

4.3. Model Verification, Testing and Inference

Comprehensive testing was performed on each model (single language, 3 languages, and 5 languages) for performance evaluations on untrained data. The data set partition was split into 70% allocation for training, 24% validation and the remaining 6% for testing, for every data set given to each model. This ensured the model was trained sufficiently, with enough data for both testing and validation. In the initialization of the model training process the Adam optimizer with a learning rate of 0.001 was used in conjunction with the Sparse Cross Entropy loss function. The models were then trained with the training set for varying epochs and batch sizes. In the validation phase the system was evaluated to gauge its overall performance. The model's accuracy, precision, recall and F1-score were calculated to provide insight on the model's capabilities. To further refine the model, a second training phase was performed. The Adam Optimizer learning rate was adjusted to 0.0001, the number of epochs and batch size were unchanged. This second training phase was designed to fine-tune the model and improve its predictive capabilities. After fine-tuning the model was evaluated again using the validation set. Once the full training process was complete, a final evaluation was performed using the test set data with the same metrics used in the validation set. In summary, the model displayed competitive performance when trained agnostically as compared to a single language model, which shows real-life applicability.

5. Experimental Evaluation and Results

The study involves three distinct experiments each serving a different purpose. The training of those are explained in detail in Section. The first experiment 5.1 compares the Wav2Vec2.0 Emissions versus Mel-Spectrograms as the chosen preprocessing method in training a CNN model for gender classification. The second experiment 2 focuses on training the model agnostically, with four different combinations of three languages for each model and another that was trained on all five languages. The goal was comparing classification performances of models that were trained agnostically and determining if those training methods can provide competitive results as opposed to language specific models. The third experiment 3 aimed to increase robustness of the classification system by using augmented data in the training process. Trance music was overlapped with multiple data segments to simulate noisy scenarios. Two models were trained in this experiment, one using a clean dataset then tested with augmented data, and the other trained with some augmented data as well.

The performance of the models trained in this research were evaluated using several indicators, accuracy, precision, recall and F1 score. These metrics 5.2 were employed due to the imbalance of the dataset.

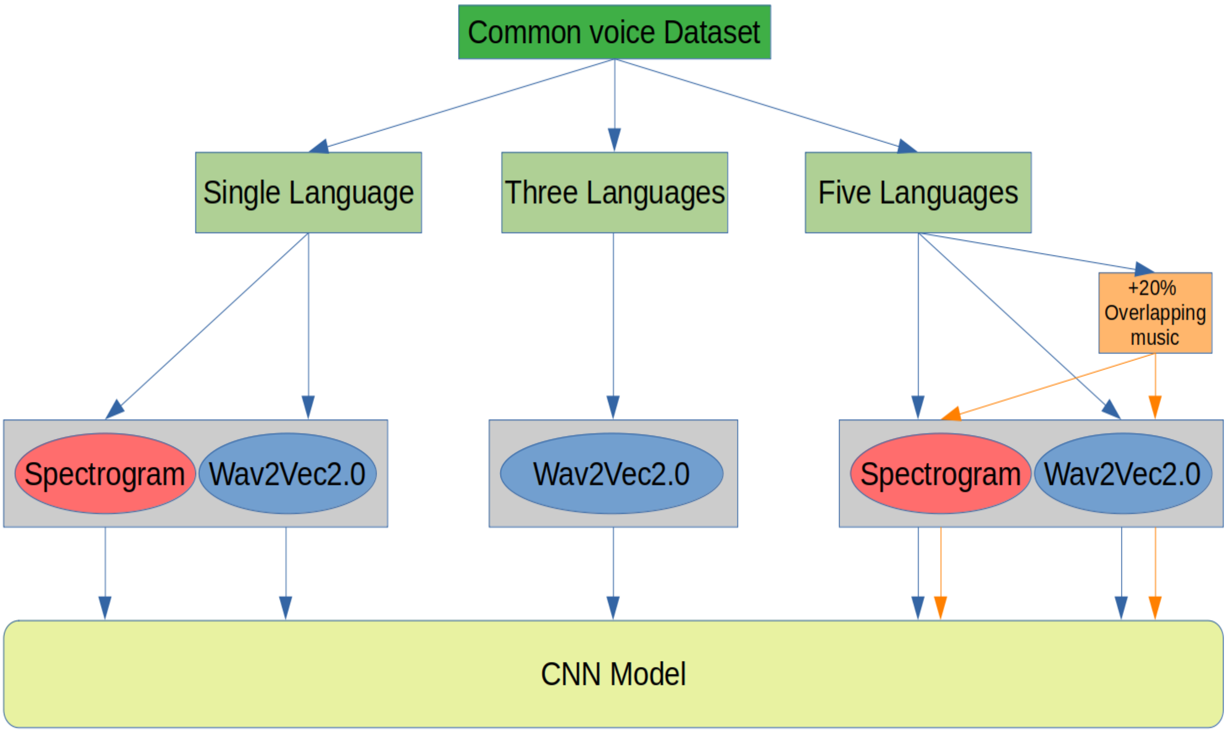

Figure 3. The figure presents the different model's architectures used in this research. In the first experiment conducted, two models were created for each language. One model was trained using the Wav2Vec2.0 emissions and the second using Mel-Spectrograms. The second experiment examines the performance of a model trained on three different languages against a model trained with five, showcasing the models generalization across languages and it's capability to work with unseen languages. The third experiment investigates the impact of augmented data on building a more robust system, comparing again Wav2Vec2.0 emissions against Mel-Spectrograms.

In the first experiment 5.3.1 the performance of the mel-spectrograms based models consistently outperformed the Wav2Vec2.0 based models across all tests performed, exhibiting higher metrics in every field. In the second experiment 5.3.2 the model trained with all five languages achieved the best performance. The models trained with different sets of three different languages showed slightly worse performance, showing that training the model agnostically results in improved robustness of the model. The third experiment 5.3.3 shows that the mel-spectrograms based model still outperforms the Wav2Vec2.0 based, across all metrics, but also shows that training the model with augmented data increases performance in noisy environments.

In the discussion section 5.4 the research concludes that using mel-spectrograms for training and testing proves better results across all languages as compared to Wav2Vec2.0 emissions. The second experiment proves that a model trained agnostically on all five languages will provide a better performing system. The third experiment shows that training the model with augmented data will provide a more robust model in noisy scenarios. Mel-spectrograms still outperformed the Wav2Vec2.0 based models but it provides better run-time in both training and testing.

5.1. Model Training

The model was trained on 3 different experiments, with different goals in mind:

1. The first experiment was performed to test the effectiveness of the Wav2Vec2.0 emissions when training a CNN model as opposed to another common preprocessing method that is converting the audio signals into a Spectrogram. In the world of speech recognition the method used in the preprocessing stage plays a vital part in the training of the model. Spectrograms represent the time frequency of the speech segment, which allows deep analysis of temporal and spectral characteristics that are inherent in human speech. By using spectrograms and Wav2Vec2.0 emission as feature extraction techniques, features can be extracted from the spectrogram and the Wav2Vec2.0 matrix, the audio signals can be reduced to a more manageable 2D representation and the usage of image processing methods can be applied.
   Two models were trained, first using spectrograms of a single language data set and second using the same data set but with Wav2Vec2.0 emissions for the training.
2. The second experiment was performed to prove that when training the model on three languages results in competitive results that are not subpar to results of a model that was trained on all five languages.
   The second experiment the model was trained language agnostically in the following way:

Trained	Tested
English, Spanish, French	Arabic, Russian
English, Spanish, Arabic	French, Russian
English, Spanish, Russian	Arabic, French
Spanish, Russian, Arabic	English, French
French, Spanish, English	Russian, Arabic

Table 3. The table above presents the data allocation regarding training and testing for each of the models in the second experiment.

   Lastly another model was trained on all five languages.
3. The last experiment goal was to train and test that if the training was done with augmented data that would help build a more robust system that performs better in noisier scenarios. The augmentation was done with overlapping trance music over data segments.
   The third experiment was split into two models as well. The first model was trained on the entire clean dataset, then tested on 20% of that data but augmented with overlapping trance music.
   The second model was trained using the entire clean dataset and 20% of that dataset but augmented, overall it was trained with 120% data. It was tested on 100% of the dataset when 20% of that was augmented.

5.2. Evaluation of the Developed Predictive Model

There are several important indicators that can be used to evaluate the performance of the model and they are: Accuracy, precision, recall, and F1 score Uddin et al. (2022). Each of them has an important role that contributes to the overall assessment. The model was provided with an unbalanced data set, the data set contains audio segments divided into five different languages English, Arabic, Spanish, French and Russia Each language contains a different number of audio recordings and, in addition, there is no equal ratio between male and female audio recordings. An unbalanced data set can be a challenge in analyzing the model evaluation and the accuracy index alone will not be enough, therefore it is necessary to evaluate the performance of the model with all the relevant indices.

Markings:

- TP = True Positive.
- FP = False Positive.
- TN = True Negative.
- FN = False Negative.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

The accuracy is used as a statistical measure and it measures the overall correctness of the model's prediction In an unbalanced data set like in our case, the accuracy index may provide a wrong estimate because it is affected by the class that contains more data, so we will combine additional indices in order to get a more reliable estimate.

$$Precision = \frac{TP}{TP + FP}$$

Precision measures the proportion of correctly predicted positive cases out of all cases predicted as positive while the model avoids false results. Through this calculation we can get a reliable situation picture since even in the case of an unbalanced data set like in our case we can correctly identify positive cases.

$$Recall = \frac{TP}{TP + FN}$$

The recall is the proportion of correctly predicted positive instances out of all the actual positive instances. It should be 1 for a good classifier. It is important because it measures the ability of the model to find and identify all the positive instances, irrespective of the class imbalance. Because of that, it provides a reliable measure of performance in an unbalanced dataset.

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

The F1-Score is a metric that combines precision and recall to provide a balanced evaluation of the model's performance it is useful when you want to consider both of them simultaneously. When the dataset is unbalanced, the accuracy can have a wrong value because of the unbalanced classes, and in these situations the F1-Score becomes valuable. The F1-Score calculates the harmonic mean of precision and recall for balance between the two. The result value represents the overall effectiveness of the model and by using it we can ensure that the model is performing well.

5.3. Experiments and Results

In this body of work, an engaging series of experiments is undertaken with a focus on advancing the understanding of deep learning techniques and their efficacy in addressing Natural Language Processing (NLP) challenges. The first experiment (Section5.3.1) embarks on an intriguing comparison of two influential techniques in the field—Transformers and Spectrogram data methods. Models are trained with both data forms, derived from identical audio recordings, aiming to foster a rigorous, direct comparison that could yield valuable insights into the strengths and shortcomings of each approach.

Subsequent experiments continue delving into complex, yet fascinating aspects of language processing and audio recognition. Experiment two (Section5.3.2) embarks on a detailed analysis of triples of languages, examining their agnostic properties against a backdrop of a five-language model. The goal here is to better understand the dynamics and interaction among languages within multilingual systems. Lastly, the third experiment (Section5.3.3) takes on a real-world challenge: the task of discerning the gender of an individual based on an audio recording, even amidst the bustling noise of environments like nightclubs or stadiums. Through these studies, new knowledge is hoped to be uncovered and more effective strategies developed for solving complex problems within the realm of NLP.

5.3.1. (Transformer+CNN) Vs. (Spectrogram+CNN)

The comparative study detailed in this section was designed to elucidate the relative performance of two combinations: Transformer data processed through a CNN and Spectrogram data also processed with a CNN. By employing these two different types of input data while maintaining the CNN as a constant, a clear picture of their respective strengths and limitations in processing different languages was hoped to be achieved.

In this experiment, a total of ten separate models were trained, one for each language in the dataset. The languages were carefully chosen to represent a diverse set of syntactical structures and phonetic characteristics, ensuring a comprehensive evaluation of the models' capabilities. This rigorous approach offered an opportunity to not only evaluate the effectiveness of Transformer and Spectrogram data but also gauge how these methods performed across a spectrum of languages, each with its unique set of challenges.

Comparison between all 5 languages using Transformer method (wac2vec)					
#	Language	Accuracy	Precision	Recall	F1 Score
1	Arabic	89.00	89.12	88.70	88.86
2	Spanish	80.24	80.42	80.22	80.21
3	English	79.33	79.33	79.33	79.33
4	Russian	**89.86**	89.85	89.87	89.86
5	French	81.14	81.15	81.13	81.14

Table 4. Performance evaluation between Wav2Vec's transformer method on our 5 sample languages. The results show mediocre performance in accuracy.

Table 4 lists the results of models that were trained on Wav2Vec's transformer emission data. The next table 5 lists the results of the same models, that were trained on the traditional spectrogram data.

Comparison between all 5 languages using Spectrogram method					
#	Language	Accuracy	Precision	Recall	F1 Score
6	Arabic	97.22	97.31	97.10	97.19
7	Spanish	97.08	97.08	97.08	97.08
8	English	93.75	93.75	93.76	93.75
9	Russian	**99.24**	99.23	99.24	99.24
10	French	97.31	97.31	97.31	97.31

Table 5. Performance evaluation between a traditional Spectrogram method. Clearly, the results show a striking disparity between Transformer and Spectrogram data. Transformer's results was consistently outshone by Spectrogram data.

The results underscore a striking disparity between Transformer and Spectrogram data techniques across the languages tested. While Transformer models performed with reasonable success, their performance was consistently outshone by Spectrogram models. Notably, Russian spectrogram model demonstrated almost perfect accuracy, precision, recall, and F1 scores (entry #9), whereas its Transformer counterpart achieved only 89.86% accuracy. (entry #4)

Choosing the most suitable model for a use case, however, may also necessitate a consideration of additional factors such as runtime. For instance, the more complex architectures, like the spectrogram, could potentially require more time for both training and testing, which may not always be feasible depending on the constraints of the use case. It is therefore advisable to take these potential trade-offs into account when selecting the most appropriate model.

5.3.2. Competitive Triples Vs. Quinta

This carefully designed experiment revolves around the exploration of language agnosticism, utilizing five distinct models trained on different combinations of five languages—English (E), Spanish (S), Russian (R), Arabic (A), and French (F). Each of the first four models is trained on a select set of three languages from the language pool. The fifth model, however, is constructed using all five languages, enabling an encompassing examination of a multilingual system. Table 6 and 7provides a detailed account of the experiment configuration. Each row represents a unique model, and the 'Train Languages' column indicates the specific languages used for training, while the 'Test Languages' column indicates the languages used to test the model and provide the scores.

By crafting this experiment, the objective is to better understand how language agnosticism operates within multilingual systems, and how different languages' interactions can impact the model's effectiveness. The diverse language combinations help to evaluate the influences of individual languages and their collective dynamics within multilingual models. The aim is to uncover insights that could guide the development and optimization of future multi-language systems.

Result comparison between different model train data, using Wav2Vec emission						
#	Train Languages	Test Languages	Accuracy	Precision	Recall	F1 Score
11	E, S, F	A, R	77.40	77.77	77.17	77.20
12	E, S, A	F, R	80.89	81.09	80.94	80.87
13	E, S, R	A, F	77.70	77.96	77.63	77.62
14	S, R, A	E, F	76.26	76.73	76.18	76.11
15	E, S, F, A, R	E, S, F, A, R	**82.82**	82.82	82.82	82.82

Table 6. Performance evaluation of various train-test language permutations, showing accuracy, precision, recall, and F1 scores for each configuration. We can see better results for the model that was trained on all languages, providing better language agnostic results.

Entry #15 shows that the language agnostic model that was trained on all 5 languages, performed the best out of all of the five. The next table 7 shows the same 5 language model, the only difference was that the data was created using traditional spectrogram method.

Result using Spectrogram data						
#	Train Languages	Test Languages	Accuracy	Precision	Recall	F1 Score
16	E, S, F, A, R	E, S, F, A, R	**88.61**	88.66	88.59	88.60

Table 7. This table shows the results for the language agnostic model that was trained on Spectrogram data. This shows worse results in comparison to 5 but still better accuracy than Wav2Vec's agnostic models (entry #15 6).

The experiments in language agnosticism revealed insightful findings. The highest performance was observed when all five languages (English, Spanish, French, Arabic, Russian) were included, achieving an identical accuracy, precision, recall, and F1 score of 82.82%. Permutations excluding two languages yielded slightly lower metrics, with the lowest scores (76.26% accuracy) being observed in the model trained on Spanish, Russian, and Arabic and then tested on English and French. These results emphasize the value of multilingual training in improving the robustness of the models. Once again, the spectrogram data-based model performed better when it comes to accuracy, precision, recall, and F1 scores. In the table above, the spectrogram performance and Wav2Vec's emission performance were directly compared. Note that, as expected, the more complex Wav2Vec method required more resources and runtime in both training and testing. So, choosing a preferred method may depend on the use-case.

5.3.3. Nightclub Scenario

In this section, the complexities of a practical, real-world scenario - a noisy nightclub environment - are navigated. This particular experiment was structured around two models. The first model was trained on data representing noisy environments, imitating the conditions of a bustling nightclub. Subsequently, this model was tested on data from both silent and noisy environments, evaluating how well it generalizes. The second model in this experiment was taken from the second experiment, a generalized model that was trained on silent data from all participating languages. In this experiment, this model is tested on noisy data, to test how well it would cope with those conditions. The two models were structured in this manner to understand how the variety in training data impacts model performance and how well a model trained in a specific environment can adapt to different ones.

The 'Train/Test Environment' section in Table 8 and 9 provides a snapshot of the unique characteristics of each model. This allows for scrutiny of how training specificity affects model performance and adaptability across diverse auditory environments. Through these comparative studies, the aim is to illuminate the challenges and opportunities in creating models that can effectively handle real-world scenarios with varying noise levels.

Result comparison between model train data, using Wav2Vec emission data						
#	Train Environment	Test Environment	Accuracy	Precision	Recall	F1 Score
17	Noisy and silent	Noisy and silent	81.19	81.19	81.19	81.19
18	Silent	Noisy	75.99	77.14	75.87	75.67

Table 8. The table shows the difference between a model that was trained on a silent environment in comparison to a model that was trained on both silent and noisy environment, according to the use-case of this section: Identifying gender by voice in a nightclub. Clearly, the model that was designated to noisy environments perform better.

Table 8 shows the importance of training the model according to it's use-case. The models shown were trained on Wav2Vec transformer emission data. The next table 9 shows the same experiment, on models that were trained on spectrogram data.

Results using Spectrogram data						
#	Train Environment	Test Environment	Accuracy	Precision	Recall	F1 Score
19	Noisy and silent	Noisy and silent	98.79	98.79	98.79	98.79
20	Silent	Noisy	88.69	90.69	88.82	88.57

Table 9. When training the model on spectrogram data, we still got the same behavior as in table 8, the model that was trained on noisy data performed better. In addition, we also see better accuracy using spectrogram data versus using Wav2Vec's transformer.

The model trained in the noisy environment was also tested in both silent and noisy environments to examine its generalizability. Using both Wav2Vec emission data and Spectrogram data, the model trained in the noisy and silent environment consistently outperformed the one trained only in the silent environment. Specifically, when using Wav2Vec emission data, the former model achieved an accuracy, precision, recall, and F1 score of 81.19, as opposed to the latter model's corresponding scores of 75.99, 77.14, 75.87, and 75.67 respectively. The performance difference was even more pronounced with Spectrogram data, with the first model scoring 98.79 across all metrics, while the second model managed 88.69 in accuracy, 90.69 in precision, 88.82 in recall, and 88.57 as the F1 score. These findings underscore the importance of training deep learning models on diverse and representative data, highlighting that a model trained in a specific environment can demonstrate superior performance when it matches the conditions it was trained under.

5.4. Discussion

When examining the outcomes of the three experiments, it becomes clear that models tested with data from the same source generally perform better. For instance, in the second experiment, a model trained on all five languages demonstrated better accuracy than models trained only on three languages. However, the difference wasn't substantial. This finding suggests the possibility of using these models on languages they weren't originally designed for, which can be beneficial when resources like time and models are limited. Additionally, in the third experiment, it was found that a model trained in a noisy environment demonstrated better accuracy compared to another model that wasn't trained with noisy recordings. Training the model in noisy environments significantly improved its gender recognition abilities, enabling it to distinguish between male and female voices more accurately. This has promising implications for voice recognition and speech analysis applications.

A direct comparison can be drawn between the five-language model that employs wav2vec's emission (entry #5) and the model designed for noisy environments (entry #17). Upon analyzing these results, it becomes evident that the accuracy doesn't significantly decline (82.82 vs 81.19) when the model is also trained in noisy environments. Another intriguing comparison can be drawn between the five-language model and each of the single language model experiments in the first experiment. The data suggests that the best performing single-language wav2vec models, specifically those for Russian and Arabic, outperformed the five-language model. Therefore, when dealing with Russian and Arabic, the preference might be towards utilizing a single-language model. Conversely, for French, English, and Spanish, the five-language model could be used as it yields comparable results. Thus, for specific languages, there is the option to utilize a more efficient, generic model.

Throughout all scenarios and experiments, a clear gap in accuracy is observed between the traditional spectrogram method and Wav2Vec's transformer method. The spectrogram method has taken a clear win in terms of accuracy, while Wav2Vec's runtime performance in both training and testing stages proves superior.

6. Conclusions and Future Work

The findings from these comparative studies present compelling evidence for the utility of both traditional spectrogram and transformer-based (Wav2Vec) methodologies in handling different languages and acoustic environments. The spectrogram method consistently demonstrated higher accuracy across various experiments, making it a robust technique for speech recognition tasks. On the other hand, the transformer approach provided superior runtime performance, emphasizing its practicality for real-time applications. This balance of accuracy and efficiency makes it a viable alternative for specific use-cases, especially when dealing with resource and time constraints.

The conducted experiments also underlined the importance of multilingual training and the model's adaptability to different environmental noise. Particularly, models trained on a diverse set of languages performed better on language-agnostic tasks, and models designed for noisy environments showed better performance under such conditions. This highlights the necessity to incorporate more diverse data and scenarios during model training to improve their robustness and versatility.

While these studies have yielded valuable insights, there are several exciting opportunities for further exploration. One of the areas of interest could be to evaluate the performance of these methods on a wider range of languages, dialects, and accents to fully appreciate their strengths and limitations. It would also be worthwhile to study the impact of different acoustic environments on model performance. By integrating more environmental conditions during the model training, it is possible to make these models even more adaptable to real-world scenarios.

Furthermore, as technologies continue to evolve, newer methodologies could be introduced. Exploring these advanced techniques, and comparing them with the existing ones, would further push the boundaries of what is achievable in the realm of speech recognition. Continuous learning and innovation in this field will no doubt lead to even more sophisticated and efficient models capable of understanding and interpreting human speech in a way that was not thought possible before.

7. Acknowledgments

This work was supported by the Ariel Cyber Innovation Center in conjunction with the Israel National Cyber directorate in the Prime Minister's Office, and Ariel Data Science and Artificial Intelligence Research Center.

References

Agarwal, S., & Chowdary, C. R. (2020). A-stacking and a-bagging: Adaptive versions of ensemble learning algorithms for spoof fingerprint detection. *Expert Systems with Applications*, *146*, 113160. URL: https://www.sciencedirect.com/science/article/pii/S0957417419308784. doi:https://doi.org/10.1016/j.eswa.2019.113160.

Alcaraz, J., Labbé, M., & Landete, M. (2022). Support vector machine with feature selection: A multiobjective approach. *Expert Systems with Applications*, *204*, 117485. URL: https://www.sciencedirect.com/science/article/pii/S0957417422008144. doi:https://doi.org/10.1016/j.eswa.2022.117485.

Alkhammash, E. H., Hadjouni, M., & Elshewey, A. M. (2022). A hybrid ensemble stacking model for gender voice recognition approach. *Electronics*, *11*. URL: https://www.mdpi.com/2079-9292/11/11/1750. doi:10.3390/electronics11111750.

Alnuaim, A. A., Zakariah, M., Shashidhar, C., Hatamleh, W. A., Tarazi, H., Shukla, P. K., Ratna, R., & Hashmi, M. F. (2022). Speaker gender recognition based on deep neural networks and resnet50. *Wirel. Commun. Mob. Comput.*, *2022*. URL: https://doi.org/10.1155/2022/4444388. doi:10.1155/2022/4444388.

Anidjar, O. H., Estève, Y., Hajaj, C., Dvir, A., & Lapidot, I. (2023). Speech and multilingual natural language framework for speaker change detection and diarization. *Expert Systems with Applications*, *213*, 119238.

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., & Weber, G. (2019). Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*, .

Arias del Campo, F., Guevara Neri, M. C., Vergara Villegas, O. O., Cruz Sánchez, V. G., de Jesús Ochoa Domínguez, H., & García Jiménez, V. (2021). Auto-adaptive multilayer perceptron for univariate time series classification. *Expert Systems with Applications*, *181*, 115147. URL: `https://www.sciencedirect.com/science/article/pii/S0957417421005881`. doi:`https://doi.org/10.1016/j.eswa.2021.115147`.

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, *33*, 12449–12460.

Beyerlein, P. (1997). Discriminative model combination. In *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings* (pp. 238–245). IEEE.

Cannon, D. F., Ferreira, R. R., & Ross, L. E. (1998). An analysis of sexual harassment in private clubs. *Journal of Hospitality & Tourism Education*, *10*, 63–71.

Chachadi, K., & Nirmala, S. (2022). Voice-based gender recognition using neural network. In *Information and Communication Technology for Competitive Strategies (ICTCS 2020) ICT: Applications and Social Interfaces* (pp. 741–749). Springer.

Choi, K., Lim, J. S., & Kim, S. (2022). Self-supervised inter-and intra-slice correlation learning for low-dose ct image restoration without ground truth. *Expert Systems with Applications*, *209*, 118072.

Cowley, A. E., & Kautzsch, E. (1910). *Gesenius' Hebrew Grammar*. Clarendon Press Oxford.

Deschamps-Berger, T., Lamel, L., & Devillers, L. (2022). Investigating transformer encoders and fusion strategies for speech emotion recognition in emergency call center conversations. In *Companion Publication of the 2022 International Conference on Multimodal Interaction* (pp. 144–153).

DeVito, J. A., O'Rourke, S., & O'Neill, L. (2000). *Human communication*. Longman New York.

Dogan, A., & Birant, D. (2021). Machine learning and data mining in manufacturing. *Expert Systems with Applications*, *166*, 114060. URL: `https://www.sciencedirect.com/science/article/pii/S095741742030823X`. doi:`https://doi.org/10.1016/j.eswa.2020.114060`.

Farghaly, A., & Shaalan, K. (2010). Arabic natural language processing: Challenges and solutions, . (pp. 8(4):1–29).

Garnerin, M., Rossato, S., & Besacier, L. (2019). Gender representation in french broadcast corpora and its impact on asr performance, . (p. 3–9). URL: `https://doi.org/10.1145/3347449.3357480`. doi:`10.1145/3347449.3357480`.

Gou, J., Sun, L., Du, L., Ma, H., Xiong, T., Ou, W., & Zhan, Y. (2022). A representation coefficient-based k-nearest centroid neighbor classifier. *Expert Systems with Applications*, *194*, 116529. URL: `https://www.sciencedirect.com/science/article/pii/S0957417422000288`. doi:`https://doi.org/10.1016/j.eswa.2022.116529`.

Grolman, E., Cohen, D., Frenklach, T., Shabtai, A., & Puzis, R. (2022). How and when to stop the co-training process. *Expert Systems with Applications*, *187*, 115841. URL: `https://www.sciencedirect.com/science/article/pii/S0957417421012021`. doi:`https://doi.org/10.1016/j.eswa.2021.115841`.

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J. et al. (2018). Recent advances in convolutional neural networks. *Pattern recognition*, *77*, 354–377.

Gupta, S. S., Hossain, S., & Kim, K.-D. (2022). Recognize the surrounding: Development and evaluation of convolutional deep networks using gammatone spectrograms and raw audio signals. *Expert Systems with Applications*, *200*, 116998.

Haeb-Umbach, R., & Ney, H. (1992). Linear discriminant analysis for improved large vocabulary continuous speech recognition. In *icassp* (pp. 13–16). Citeseer volume 92.

Hendrycks, D., Mazeika, M., Kadavath, S., & Song, D. (2019). Using self-supervised learning can improve model robustness and uncertainty. *Advances in neural information processing systems*, *32*.

Hord, L. C. (2016). Bucking the linguistic binary: Gender neutral language in english, swedish, french, and german. *Western Papers in Linguistics*, *3*.

Huang, X., Li, Z., Jin, Y., & Zhang, W. (2022). Fair-adaboost: Extending adaboost method to achieve fair classification. *Expert Systems with Applications*, *202*, 117240. URL: `https://www.sciencedirect.com/science/article/pii/S0957417422006182`. doi:`https://doi.org/10.1016/j.eswa.2022.117240`.

Jahangir, R., Teh, Y. W., Nweke, H. F., Mujtaba, G., Al-Garadi, M. A., & Ali, I. (2021). Speaker identification through artificial intelligence techniques: A comprehensive review and research challenges. *Expert Systems with Applications*, *171*, 114591.

Janeva, T., Mishev, K., & Simjanoska, M. (2022). Language agnostic voice recognition model.

Kalaycı, E. E., & Doğan, B. (2020). Gender recognition by using acoustic features of sound with deep learning and data mining methods. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)* (pp. 1–4). doi:`10.1109/ASYU50717.2020.9259824`.

Karlos, S., Kanas, V. G., Aridas, C., Fazakis, N., & Kotsiantis, S. (2019). Combining active learning with self-train algorithm for classification of multimodal problems. In *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)* (pp. 1–8). doi:`10.1109/IISA.2019.8900724`.

Key, M. R. (2011). The relationship of verbal and nonverbal communication. In *The relationship of verbal and nonverbal communication*. De Gruyter Mouton.

Lastow, F., Ekberg, E., & Nugues, P. (2022). Language-agnostic age and gender classification of voice using self-supervised pre-training. In *2022 Swedish Artificial Intelligence Society Workshop (SAIS)* (pp. 1–9). IEEE.

Livieris, I. E., Pintelas, E., & Pintelas, P. (2019). Gender recognition by voice using an improved self-labeled algorithm. *Machine Learning and Knowledge Extraction*, *1*, 492–503.

Lundberg, S. M., Erion, G. G., & Lee, S.-I. (2018). Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*, .

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

Meel, P., & Vishwakarma, D. K. (2021). A temporal ensembling based semi-supervised convnet for the detection of fake news articles. *Expert Systems with Applications*, *177*, 115002. URL: `https://www.sciencedirect.com/science/article/pii/S0957417421004437`. doi:`https://doi.org/10.1016/j.eswa.2021.115002`.

Mellgren, C., Andersson, M., & Ivert, A.-K. (2018). "it happens all the time": Women's experiences and normalization of sexual harassment in public space. *Women & Criminal Justice*, *28*, 262–281.

Moscatelli, M., Parlapiano, F., Narizzano, S., & Viggiano, G. (2020). Corporate default forecasting with machine learning. *Expert Systems with Applications*, *161*, 113567. URL: `https://www.sciencedirect.com/science/article/pii/S0957417420303912`. doi:`https://doi.org/10.1016/j.eswa.2020.113567`.

Mushava, J., & Murray, M. (2022). A novel xgboost extension for credit scoring class-imbalanced data combining a generalized extreme value link and a modified focal loss function. *Expert Systems with Applications*, *202*, 117233. URL: `https://www.sciencedirect.com/science/article/pii/S0957417422006133`. doi:`https://doi.org/10.1016/j.eswa.2022.117233`.

Nur, M. M. T., Dola, S. S., Banik, A. K., Akhter, T., Hossain, N., Al Islam, A. A., & Noor, J. (2022). Speaker identification through gender detection. In *2022 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)* (pp. 181–188). IEEE.

Özcan, Z., & Kayıkçıoğlu, T. (2021). Evaluating mfcc-based speaker identification systems with data envelopment analysis. *Expert Systems with Applications*, *168*, 114448.

Pahwa, A., & Aggarwal, G. (2016). Speech feature extraction for gender recognition. *International Journal of Image, Graphics and Signal Processing*, *8*, 17.

Perry, T. L., Ohde, R. N., & Ashmead, D. H. (2001). The acoustic bases for gender identification from children's voices. *The Journal of the Acoustical Society of America*, *109*, 2988–2998. URL: `https://doi.org/10.1121/1.1370525`. doi:`10.1121/1.1370525`. `arXiv:https://pubs.aip.org/asa/jasa/article-pdf/109/6/2988/8088947/2988_1_online.pdf`.

Pikuliak, M., Šimko, M., & Bieliková, M. (2021). Cross-lingual learning for text processing: A survey. *Expert Systems with Applications*, *165*, 113765.

Prinzie, A., & Van den Poel, D. (2008). Random forests for multiclass classification: Random multinomial logit. *Expert systems with Applications*, *34*, 1721–1732.

Shahbazi, N., Lin, Y., Asudeh, A., & Jagadish, H. V. (2023). Representation bias in data: A survey on identification and resolution techniques. *ACM Computing Surveys*, . URL: `https://doi.org/10.1145%2F3588433`. doi:`10.1145/3588433`.

Simonović, M., Kovandžić, M., Ćirić, I., & Nikolić, V. (2021). Acoustic recognition of noise-like environmental sounds by using artificial neural network. *Expert Systems with Applications*, *184*, 115484.

Submitter, I., Jena, B., Mohanty, A., Mohanty, S. K. et al. (2021). Gender recognition and classification of speech signal. In *Proceedings of the International Conference on Smart Data Intelligence (ICSMDI 2021)*.

Sun, Y., Li, J., Xu, Y., Zhang, T., & Wang, X. (2023). Deep learning versus conventional methods for missing data imputation: A review and comparative study. *Expert Systems with Applications*, (p. 120201).

Tursunov, A., Mustaqeem, Choeh, J. Y., & Kwon, S. (2021a). Age and gender recognition using a convolutional neural network with a specially designed multi-attention module through speech spectrograms. *Sensors*, *21*. URL: `https://www.mdpi.com/1424-8220/21/17/5892`. doi:`10.3390/s21175892`.

Tursunov, A., Mustaqeem, Choeh, J. Y., & Kwon, S. (2021b). Age and gender recognition using a convolutional neural network with a specially designed multi-attention module through speech spectrograms. *Sensors*, *21*, 5892.

Uddin, M. A., Pathan, R. K., Hossain, M. S., & Biswas, M. (2022). Gender and region detection from human voice using the three-layer feature extraction method with 1d cnn. *Journal of Information and Telecommunication*, *6*, 27–42.

Walker, P., McClaran, N., Zheng, Z., Saxena, N., & Gu, G. (2022). Biashacker: Voice command disruption by exploiting speaker biases in automatic speech recognition, . (p. 119–124). URL: `https://doi.org/10.1145/3507657.3528558`. doi:`10.1145/3507657.3528558`.

Wang, C., Wu, Y., Qian, Y., Kumatani, K., Liu, S., Wei, F., Zeng, M., & Huang, X. (2021). Unispeech: Unified speech representation learning with labeled and unlabeled data. In *International Conference on Machine Learning* (pp. 10937–10947). PMLR.

Wang, F., & Liu, H. (2021). Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2495–2504).

Wright, B. (2002). Gender and language: challenging the stereotypes. *Unpublished essay*, .

Wu, J.-D., & Tsai, Y.-J. (2011). Speaker identification system using empirical mode decomposition and an artificial neural network. *Expert Systems with Applications*, *38*, 6112–6117.

Xie, J., & Zhu, M. (2019). Investigation of acoustic and visual features for acoustic scene classification. *Expert Systems with Applications*, *126*, 20–29. URL: `https://www.sciencedirect.com/science/article/pii/S0957417419300661`. doi:`https://doi.org/10.1016/j.eswa.2019.01.085`.

Zhao, J., & Zhang, W.-Q. (2022). Improving automatic speech recognition performance for low-resource languages with self-supervised models. *IEEE Journal of Selected Topics in Signal Processing*, *16*, 1227–1241.

Zolnay, A., Schluter, R., & Ney, H. (2005). Acoustic feature combination for robust speech recognition. In *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.* (pp. I–457). IEEE volume 1.

Zoughi, T., Homayounpour, M. M., & Deypir, M. (2020). Adaptive windows multiple deep residual networks for speech recognition. *Expert Systems with Applications*, *139*, 112840.