Available online at www.sciencedirect.com
ScienceDirect
Expert Systems with Applications 00 (2023) 1–16
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

Or & Roi: This is our section
Amit
**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 reached better results compared to the Wav2Vec Transformer method For the Spanish language we reached a result of 100% in all the Accuracy Precision Recall F1 Score indicators while the Transformer equivalent achieved only 80.24% accuracy**

*Keywords:* Machine Learning,

## 1. Introduction

Amit & Moriah: This is your section
Moriah
It is common to say that in human communication most of the information is not verbal.
Beyond the words that someone says, there are also many details about the person who speaks, like body language, intonation, and more. Humans use variations in pitch, speed, volume, tone, rhythm, body movements, and facial expressions. It all helps convey additional information and provide a sense of the message.
When it comes to Automatic Speech Recognition (ASR) systems, the understanding that human communication has more than just words is crucial.

*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)

The words are called transcripts, on ASR the audio file is converted to speech that is, meaningful words. But there is an additional layer beyond the transcription of spoken words that is equally important, the metadata.

The metadata contains various elements such as speaker identity, speech intonation, emotions during the speech, and more. By extracting this metadata from raw audio, ASR systems will have a deeper understanding of the speech, and understanding of the conversation will improve.

The human voice contains personality traits such as: language, age, gender,Uddin et al. (2022) the unique shape of the vocal system Pahwa & Aggarwal (2016), accent, rhythm and more.

In this paper, we focus on gender classification. Gender classification by voice refers to the process of determining gender based on analyzing different parameters extracted from the voice sample, classifying it as a female voice or a male voice Submitter et al. (2021).

There are quite a few motivations for a gender classification system. 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. It is very important to create an environment for women's well-being, comfort, and security, especially in a club, where the music and loud noise make it difficult to detect that something wrong is happening. 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. It can help to provide a sense of safety and prevent potential cases of harassment or discomfort.

A significant advantage of a voice recognition system in public places is the creation of a balance between privacy and security by providing an effective identification tool that avoids the need for visual surveillance that causes privacy violations, like security cameras.

Amit

Distinguishing other people's voices may seem like a trivial task. If you are trying to talk to a person who does not speak the same language, the task becomes much more challenging. Even speakers of the same language coming from different regions will not be able to understand each other.

For example, the Arabic language is a very diverse Semitic language, containing different dialects. There are many words that are pronounced differently and in different sentence structures that can cause a communication problem and a lack of understanding between speakers. Farghaly & Shaalan (2010):

Each language has its own characteristics and components that together create its uniqueness. There are recognized languages which are "Gender Neutral" These languages are free from assumptions about the type of social gender or biological sex of people to whom they are behavioral in speech. These languages have a vocabulary and sentences that are not gender-specific, that is, they use words or sentences for both genders. For example, the English language is gender neutralHord (2016):

In this language, the use of words, phrases, and sentences allows referring to a certain person or group without specifying their gender. On the other hand, there are languages that are not gender neutral, these languages are characterized by the fact that words in the language are sexually differentiated. There are differences in the structure of the language, for example adding a certain suffix to a word can indicate whether a word belongs to a male or a female. Or for example adjectives referring to a certain gender.

For example, in the Hebrew language Cowley & Kautzsch (1910): there are masculine and feminine forms that refer to names, nouns, adjectives, and pronouns. In both cases, a major problem arises because people who are not sexually defined or people who choose for a certain reason to use words and sentences that are sexually diagnosed as a certain type of gender but use them in the opposite way. This problem can create a false distinction and therefore we cannot determine the speaker's gender the use of these words.

On the other hand, speakers of gender-neutral languages can also be a problem, because since these languages are not sexually diagnosed, we cannot always know the gender of the speaker. 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, 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.

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, 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 model. To do this, we first tried speech recognition using the wav2vec method.Baevski et al. (2020) It is a self-supervised learning 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. 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.
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.
In addition, Mel-spectrogram provides an accurate and high-quality display of the frequencies. Using this method, we can perform feature extraction in a precise and correct way, which is a necessary step in building the model.

Moriah

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 CommonVoice. 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. 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. The second experiment was performed while using Wav2Vec2 emission. The data was divided into groups so that each group has a combination of three different languages. 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 performed also on the five languages together. The third experiment was similar to the second, but this time it was meant to check the durability of the model while audio segments combined with trance music. The experiment was divided into two parts. The first was trained with 80% original data and 20% from the rest data with trance music pieces, the test was performed with 20% of original data and also data with trance music pieces. The second was trained with 100% original data and the test was performed with 20% of data with trance music pieces.

### *1.1. Our Contributions*

Daniel & Raz: This is your section

Finally, 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). We provide an insightful comparison of Transformer and Spectrogram data techniques when used in conjunction with Convolutional Neural Networks (CNN), exploring their individual strengths and weaknesses across a diverse set of languages. Additionally, we investigate language agnosticism within multilingual systems, examining the dynamics and interplay between various languages. Finally, we apply these techniques in a practical scenario, studying the impact of different auditory environments on model performance.

### *1.2. Paper Structure*

Amit & Moriah: This is your section

Amit

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 5.1, 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 5.2;
- Finally, Section 6 concludes and summarizes this paper.

For ease of reading, we provide a list of abbreviations in Table 1.
Amit & Moriah: This is your section

| **Abbreviation** | **Meaning** |
|---|---|
| ALAGD | Acoustic and Language Agnostic Gender Detection |
| ASR | automatic speech recognition |
| 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 |

Table 1. List of Abbreviations. Sort by alphabetic order

## 2. Related Work

Amit & Moriah: This is your section
Stav & Sivan & Hodaya: This is your section
Amit
Studies that contributed to the project through interesting techniques and tools Chachadi & Nirmala (2022)
focuses on identifying a speech signal according to the speaker's gender, it refers to the important and unique features that can be extracted from a speech segment while focusing on pitch and frequency.
Their 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 software the recognition can be improved. Mel Frequency Cepstral Coefficient (MFCC) is used MFCC knows how to identify the central components of a speech signal and discard unnecessary things. in its effectiveness in capturing the spectral characteristics of speech signals. Another use was the mel spectrogram. This feature shows the mel frequencies in the sound wave.
MS is obtained from the conversion of the sound wave into the frequency space and is mainly used for predicting voice quality and identifying frequency centrality in speech. In addition, a combination of MFCC and mel spectrogram was used. This combination reached the highest results for an accuracy of 94.32%. The proposed model highlights the ability of a neural network to effectively learn from features and improve gender recognition from an audio clip.
Another interesting paper by 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 article 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.

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

The study investigates the effectiveness of combining Mel-Spectrogram features with other feature representations in speech recognition tasks. The paper examined 2 main methods for combining the acoustic characteristics: LDA based combination and log-linear based combination.

In addition, they examined the combination of advanced acoustic characteristics of the MFCC type. The paper describes how the Mel-spectrogram is used as a representative design for sound signals in order to improve the performance of the identifier based on a sound model. From the results of the study it emerged that the best voice recognition appears when using a combination of Mel-spectrogram features with other features. The best combination is the result of combining MFCC and feature votes using the LDA-based method. It can be concluded from this paper that combining the right features and the appropriate methods, shows a guaranteed improvement in voice recognition in systems based on models that use Mel-spectrogram

Moriah

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, and with English TIMIT dataset, 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 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.

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 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 is better than other models in accuracy of more than 90% for all three classification tasks. They use SHAP method which is based on the concepts of game theory, in order to explain the predictions of machine learning model by calculating the contribution of each feature to the prediction. The results highlight the influence of specific features on gender and age range prediction.

In Tang et al. (2022) they use an algorithm which is used to embed the gender and nationality information into the spectrogram features and make full use of the information.

This makes a big difference for them in the speaker identification results. The use of the ATGN algorithm embeds the information about nationality and gender into the spectrogram features based on the attention mechanism and combines the generated gender and nationality high-level embeddings with spectrogram low-level embeddings for classification. They prove this algorithm is an effective method to explore the information about the gender and nationality that is shown in the dataset VoxCeleb1. They achieved an amazing result of 98.31% accuracy in gender classification and 93.57% accuracy in nationality classification.

Stav

In Livieris et al. (2019), in order to identify the speaker's gender, a semi-supervised algorithm, called iCST-Voting for the gender recognition by voice. This algorithm constitutes an ensemble of the most popular self-labeled algorithms. i.e., Self-training, Co-training and Tri training. two experiments took place in two distinct phases: In the first phase evaluate the performance of the iCST-Voting, against its component self-labeled algorithms: Self-training, Co-training and Tri-training and the state-of-the-art self-labeled algorithms: SETRED, Co-Bagging, Democratic-Co learning and Co-Fores. while in the second phase, compare the performance of the proposed algorithm iCST-Voting against 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%. The disadvantage of the article is that its data set is relatively small.

Another article Perry et al. (2001), diagnosed the acoustic characteristics of children's speech and voices that contribute to listeners' ability to identify their gender. Two experiments were conducted to collect and analyze data. In Experiment I, vocal recordings and physical measurements were obtained from children of different age groups: 4, 8, 12, and 16 years old. Approximately 10 girls and 10 boys were included in each age group. The speech samples consisted of seven vowels in American English. The researchers measured fundamental frequency (f0) and formant frequencies (F1, F2, F3) from these syllables. In Experiment II, 20 adults listened to the recorded syllables produced by the children in Experiment I. They rated 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. A disadvantage of this study stems from the fact that the audio samples were of children only.

In Alkhammash et al. (2022), a stacked ensemble for gender voice recognition model is presented, using four classifiers, namely, k-nearest neighbor (KNN), support vector machine (SVM), stochastic gradient descent (SGD), and logistic regression (LR) as base classifiers and linear discriminant analysis (LDA) as meta classifier. The research performance of the proposed model was compared with traditional machine learning models, where the proposed model achieved the best results for accuracy of 99.64%. The purpose of this article is to demonstrate how different machine learning models can be stacked. In this stacked model, there are 5 machine learning models. Four models are used as the base classifier and one model is applied as a meta classifier. Data preprocessing and k-fold cross-validation are used to obtain the best-predicted output. In this article the dataset is balanced, an equal number of recordings for female and for male, although the amount of the dataset in general is small containing only 3168 recordings.

In Kalaycı & Doğan (2020), Identifying the gender of the speaker was determined according to the acoustic characteristics of the speeches compiled from seven different languages by using the acoustic features of the voice. In addition to data mining techniques, deep learning was also used in the classification process. The article presents the use of different machine learning algorithms - Logistic regression, SVM, KNN, XGBoost and AdaBoost. The best success rate was achieved respectively with gradient boosting and random forest in classification process. Deep learning method with Multilayer Perceptron (MLP) was achieved 96.22% correct classification rate in their analysis.

The advantage 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

Sivan

The current study recognizes the importance of developing an ASR model that is trained on several languages to enable more accurate identification.Thus, the Common Voice dataset was utilized. `https://commonvoice.mozilla.org/he` This dataset is used in many studies on gender classification in recent years such as: Agarwal & Zesch (2021), 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 the 5 languages: English, Spanish, French, Russian and Arabic among the 108 languages that the data contains.The data consists of 3-second voice segments of women and men. Here is a graph that shows the percentages of recordings by men and women in the 5 languages that the study deals with.

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

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 you can see from the table, in all languages there is a higher number of recordings for men than women.Also, there is a fairly high difference between the number of recordings in the English, French and Spanish languages and the Arabic language. This problem 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 in women and men.

### *3.1. Gender Bias*

Hodaya

Gender bias in ASR 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 ASR 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 |

Table 3. 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.

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, leading to less available voice data for training ASR systems. By acknowledging and tackling these biases, we aim to promote inclusivity, fairness, and accuracy in ASR technologies across different genders and languages.

## 4. Framework

Itamar & Firas: This is your section

### *4.1. Preliminary operations.*

Itamar

The preliminary stage of the experiments conducted in this paper the data was prepared for every one of the experiments then pre processed according to the agenda of each one.

#### *4.1.1. Dataset preparation*

Itamar

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. 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. In this augmented and agnostic dataset we hoped to build a robust and practical model that is rich in real life use cases.

#### *4.1.2. Pre processing*

Itamar

In this research the usage of mel-spectrograms and Wav2Vec2.0 are present. These are methods of converting audio signals into 2D matrix representations, moving 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. This conversion allows the usage of image processing techniques and to leverage the power of a CNN model for feature extraction.

Every audio segment from the data set was represented by a 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 to obtain the frequency domain representation and then a mel-scaled filterbank is applied to the power spectrum received by the FFT. Logarithmic scaling is performed on the filterbank 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

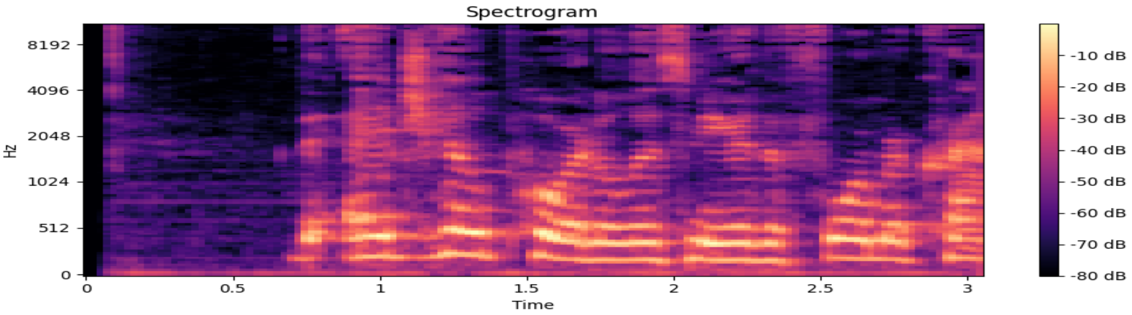

Figure 1. A example of a spectrogram that is 3 seconds long.

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.

Firas

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*

Firas

The CNN model that was used in all the conducted experiments is built with multiple convolutional layers, batch normalization and max pooling layers. The convolutional layers have varying kernel and stride sizes, which extract features from the input segment, then batch normalization is used to normalize the feature maps.After each convolution the ReLU activation function was called to prevent non-linearity. Dropouts operations are performed to prevent overfitting of the data. Layers of max-pooling are then utilized to downsample the normalized feature maps, after the last downsampling layer, a linear transformation is applied to the entire feature map collection where the resulting output is the linear transformed representation of all the extracted features. The final linear transformation results in a binary vector that predicts the gender of the speaker.

### 4.3. *Model Training*

Itamar

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.

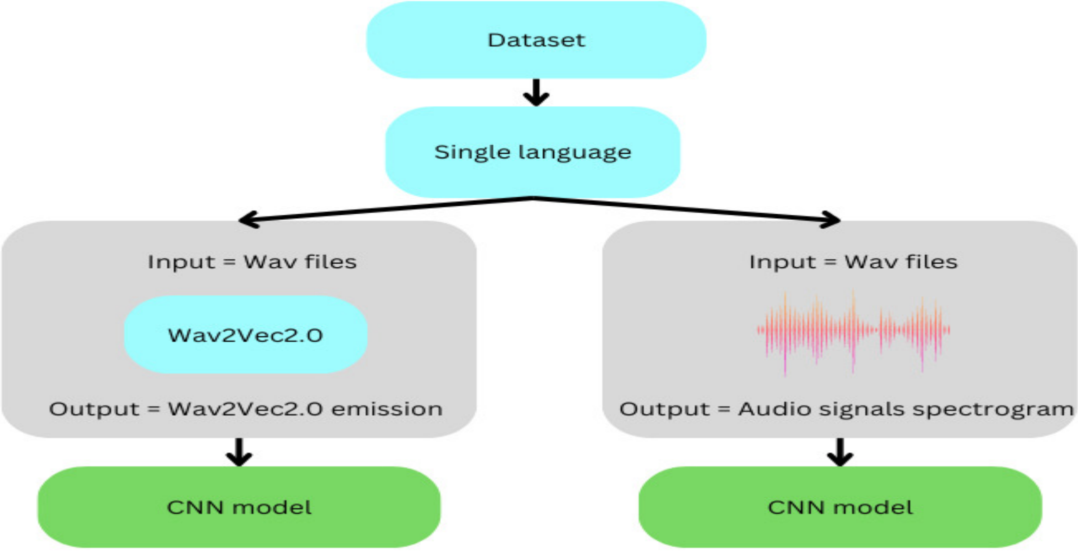

Figure 2. The first experiment conducted, two models created for each language. One model was trained using the Wav2Vec2.0 emissions and the second using Spectrograms.

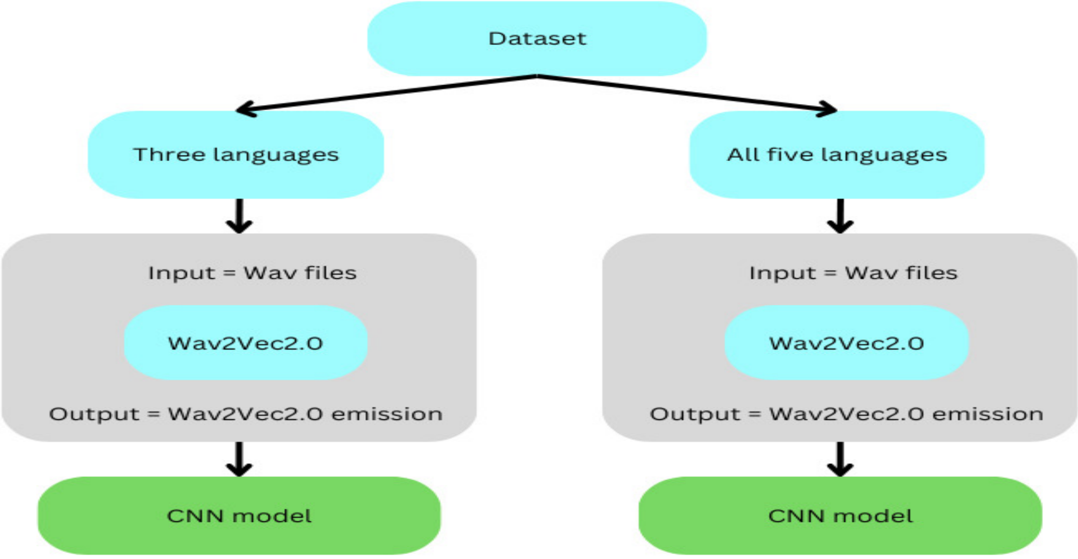

Figure 3. The second experiment conducted, five models were trained using three languages at a time and tested by the other two. A six model was trained using all five languages.

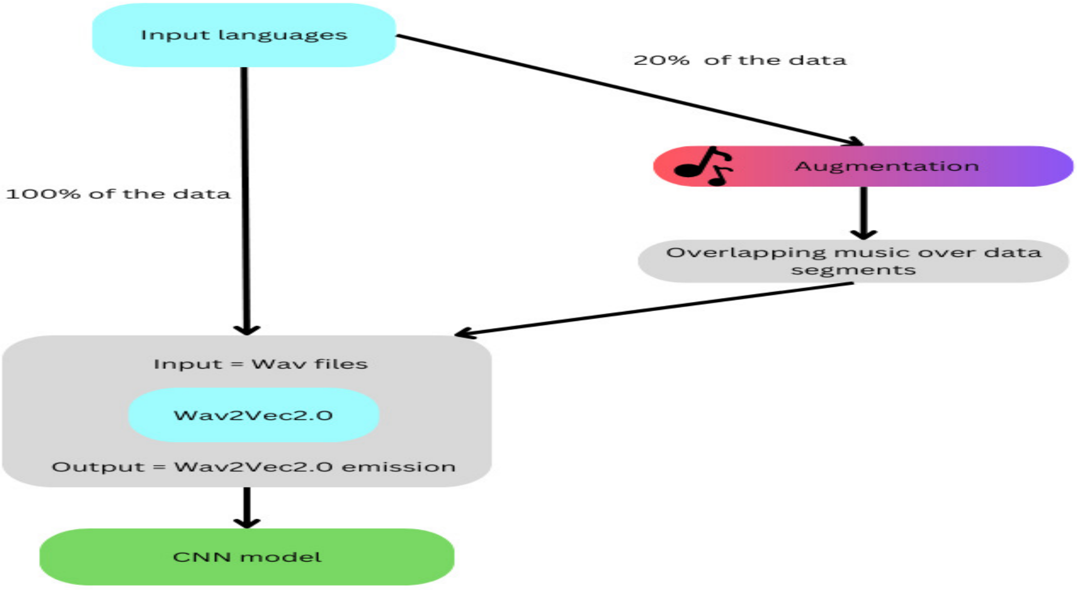

Figure 4. The third experiment conducted, the model was trained with augmented data then tested to see if the model has better performance on noisier input.

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 |

   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.

### *4.4. Model Verification, Testing and Inference*

Firas

Comprehensive testing was performed on each model for performance evaluations on untrained data. The first experiment for example the data set partition was split into 70% allocation for training, 24% validation and the remaining 6% for testing. This ensured the model was trained sufficiently, with enough data for 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 model was then trained using the training set for 5 epochs with a batch size of 10. In the validation phase the system was evaluated to gauge its overall performance. To avoid misevaluating the model since gender bias was presented in the dataset, 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 while 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

### *5.1. Evaluation of the Developed Predictive Model*

Amit & Moriah: This is your section

amit

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. 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.

Moriah

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

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.

### *5.2. Experiments and Results*

Daniel & Raz: This is your section

In this body of work, we undertake an engaging series of experiments with a focus on advancing our grasp of deep learning techniques and their efficacy in addressing Natural Language Processing (NLP) challenges. The first experiment (Section5.2.1) embarks on an intriguing comparison of two influential techniques in the field—Transformers and Spectrogram data methods. By training models with both data forms, derived from identical audio recordings, we aim to foster a rigorous, direct comparison that could yield valuable insights into the strengths and shortcomings of each approach.

The subsequent experiments continue to delve into complex, yet fascinating aspects of language processing and audio recognition. Experiment two (Section5.2.2) 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 understand better the dynamics and interaction among languages within multilingual systems. Lastly, the third experiment (Section5.2.3) 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, we hope to uncover new knowledge and develop more effective strategies for solving complex problems within the realm of NLP.

#### *5.2.1. (Transformer+CNN) Vs. (Spectrogram+CNN)*

Daniel & Raz: This is your section The comparative study detailed in this section aimed to elucidate the relative performance of two combinations: Transformer data processed through a Convolutional Neural Network (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, we hoped to achieve a clear picture of their respective strengths and limitations in processing different languages.

In this endeavor, we trained a total of ten separate models, one for each language in our 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 us 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 - Transformer | 89.00 | 89.12 | 88.70 | 88.86 |
| 2 | Spanish - Transformer | 80.24 | 80.42 | 80.22 | 80.21 |
| 3 | English - Transformer | 79.33 | 79.33 | 79.33 | 79.33 |
| 4 | Russian - Transformer | **89.86** | 89.85 | 89.87 | 89.86 |
| 5 | French - Transformer | 81.14 | 81.15 | 81.13 | 81.14 |
| Comparison between all 5 languages using Spectrogram method | | | | | |
| # | Language | Accuracy | Precision | Recall | F1 Score |
| 1 | Arabic - Spectrogram | 97.22 | 97.31 | 97.10 | 97.19 |
| 2 | Spanish - Spectrogram | **100.0** | 100.0 | 100.0 | 100.0 |
| 3 | English - Spectrogram | 93.75 | 93.75 | 93.76 | 93.75 |
| 4 | Russian - Spectrogram | 99.24 | 99.23 | 99.24 | 99.24 |
| 5 | French - Spectrogram | —- | —- | —- | —- |

Table 4. Performence evaluation between a traditional Spectrogram method versus Wav2Vec Transformer method. Clearly, the results show a striking disparity between Transformer and Spectrogram data. Transformer's results was consistently outshone by Specttrogram data.

Our results underscore a striking disparity between Transformer and Spectrogram data techniques across the languages we tested. While Transformer models performed with reasonable success, their performance was consistently outshone by Spectrogram models. Spectacularly, Spanish demonstrated perfect accuracy, precision, recall, and F1 scores when analyzed using Spectrogram data, whereas its Transformer counterpart achieved only 80.24% accuracy.

### *5.2.2. Competitive Triples Vs. Quinta*

Daniel & Raz: This is your section Our 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 our language pool. The fifth model, however, is constructed using all five languages, enabling an encompassing examination of a multilingual system.

Table 6 provides a detailed account of our experiment configuration. Each row represents a unique model, and the 'Train Languages' column indicates the specific languages used for training, and the 'Test Languages' column indicates the languages used to test this model and provide the scores.

By crafting this experiment, our objective is to understand better 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. We aim to uncover insights that could guide the development and optimization of future multi-language systems.

| Comparison between all experimented permutations | | | | | | |
|---|---|---|---|---|---|---|
| # | Train Languages | Test Languages | Accuracy | Precision | Recall | F1 Score |
| 1 | E, S, F | A, R | 77.40 | 77.77 | 77.17 | 77.20 |
| 2 | E, S, A | F, R | 80.89 | 81.09 | 80.94 | 80.87 |
| 3 | E, S, R | A, F | 77.70 | 77.96 | 77.63 | 77.62 |
| 4 | S, R, A | E, F | 76.26 | 76.73 | 76.18 | 76.11 |
| 5 | E, S, F, A, R | E, S, F, A, R | **82.82** | 82.82 | 82.82 | 82.82 |

Table 5. Performance evaluation of various train-test language permutations, showing accuracy, precision, recall, and F1 scores for each configuration. We can clearly see consistantly better results for the model that was trained on all languages, providing better langauge agnostic results.

Our experiments in language agnosticness also 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 our models.

### *5.2.3. Nightclub Scenario*

Daniel & Raz: This is your section In this section, we navigate the complexities of a practical, real-world scenario - a noisy nightclub environment. This particular experiment was structured around two models. The first model was trained exclusively on data representing noisy environments, imitating the conditions of a bustling nightclub. Subsequently, this model was tested on data from both quiet and noisy environments, evaluating how well it generalizes beyond its training conditions.

The second model in this experiment was trained on data from a broader range of environments, both quiet and noisy. Following training, this model was also tested on both quiet and noisy environments. 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 'Architectural Details' section in Table 6 provides a snapshot of the unique characteristics of each model. This allows us to scrutinize how training specificity affects model performance and adaptability across diverse auditory environments. Through these comparative studies, we aim to illuminate the challenges and opportunities in creating models that can effectively handle real-world scenarios with varying noise levels.

| Comparison between all experimented permutations | | | | | |
|---|---|---|---|---|---|
| # | Architectural Details | Accuracy | Precision | Recall | F1 Score |
| 1 | Trained on noisy environment only and tested on both environment | **8**1.19 | 81.19 | 81.19 | 81.19 |
| 2 | Trained on all environment only and tested on both environment | -— | —- | —- | —- |

Table 6. Comparision of our usecase - noisy environment

### *5.3. Discussion*

Daniel & Raz: This is your section

The predicted BD curve has been produced for each medium of the available collection of $N$ measured porous media, partially presented in ..

- The restriction of $\omega_{p_i} \geq 0$ is imposed ..
- Shortage of data; wide scattering of wetting and drying curve pairs around the actual regularity.

In addition, there are three other possible research directions to be thoroughly considered in a subsequent study:

1. Use of the non-optimized functions $S^{\circ}_{d_i}(\psi)$ for approximated linear combinations.
2. Use of the approximated linear combinations of..

## 6. Conclusions and Future Work

Or & Roi: This is your section

After the publication of the Mualem (1977, 1984); Mualem & Beriozkin (2009) dependent domain theory of capillary hysteresis, i.e. more than three decades, the transition from the HD curve to the BD curve remained a missing link in the hysteresis modeling based on this theory. This is explained probably by the significant difficulty of accounting for the spatial arrangement of the hysterons, that causes the blockage phenomenon. The suggested model is aimed to fill up this missing link. Prediction of the desired BD curve from its associated known BW curve is obtained as a product of two mappings: (i) a nonlinear mapping of the known BW curve to its corresponding HD curve and (ii) a linear mapping of this latter curve to the desired BD curve, by the suggested algorithm HD-opt-BD. We discern two approaches to predictive modeling of the inter-functional bijective relationships between the porous media characteristic functions: (i); the integral operator approach, as presented by Eq. (12), and (ii) the direct transformation as implemented in the machine learning algorithm HD-opt-BD, suggested in this paper.

## 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, A., & Zesch, T. (2021). Robustness of end-to-end automatic speech recognition models - A case study using mozilla deepspeech. *CoRR*, *abs/2105.09742*. URL: https://arxiv.org/abs/2105.09742. arXiv:2105.09742.

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.

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*, .

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.

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.

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

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.

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*.

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.

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.

Mualem, Y. (1977). Extension of the similarity hypothesis used for modeling the soil water characteristics. *Water Resources Research*, *13*, 773–780.

Mualem, Y. (1984). A modified dependent-domain theory of hysteresis. *Soil science*, *137*, 283–291.

Mualem, Y., & Beriozkin, A. (2009). General scaling rules of the hysteretic water retention function based on mualem's domain theory. *European journal of soil science*, *60*, 652–661.

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.

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.

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)*.

Tang, Y., Liu, C., Leng, Y., Zhao, W., Sun, J., Sun, C., Wang, R., Yuan, Q., Li, D., & Xu, H. (2022). Attention based gender and nationality information exploration for speaker identification. *Digital Signal Processing*, *123*, 103449.

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, 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).

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.