Voice Conversion using Classical Machine Learning Methods

Jivitesh Sabharwal
Doctor of Philosophy
PhD/CSE
jiviteshs@iiitd.ac.in

Vishesh Jain
Bachelor of Technology
BTech/CSB
vishesh20550@iiitd.ac.in

Harsh Vardhan Singh
BTech/ECE
Bachelor of Technology
harsh20202@iiitd.ac.in

Siddhant Singh
Bachelor of Technology
BTech/CSAM
siddhant20338@iiitd.ac.in

Aekansh Kathunia
Bachelor of Technology
BTech/ECE
aekansh21127@iiitd.ac.in

Abstract

This project explores voice conversion using classical machine-learning techniques applied to the Common Voice: A Massively-Multilingual Speech Corpus data set [1]. We aim to uncover insights into voice characteristics through an extensive exploratory data analysis (EDA) and an examination of existing analysis. Our project report will provide a comprehensive overview of our findings, inferences, and the novel approaches we plan to implement for voice conversion.

1 Introduction

Voice Conversion is a task in the audio data-related domain in which the audio of a specific source speaker's voice is transformed or converted into the target speaker's voice without changing the linguistic contents of the sentence spoken by the source speaker with the help of artificial intelligence. It is one of the most dominant research domains in audio signal processing. In such type of voice conversion, we focus on modifying the speaker-dependent speech features such as the supra-segmental features and segmental features[21].

Supra-segmental features, also known as the prosodic features, are speech features that are pitch, intonation, speaking speed(tempo), stress on the words and loudness, whereas the segmental features compromises of multiple phonemes that are present in spoken languages and may vary from one spoken language to another spoken language, for example, there are 44 different phonetic sounds in English, whereas French has around 37 different phonetic sounds.

One can categorize voice conversion methods based on multiple factors, which are 1) *Parallel* or *Non-Parallel* audio recordings, 2) *Text-dependent* or *Text-independent*, 3) *Language-dependent* or *Cross Language Conversion* [14]. *Parallel* audio recordings refer to the recordings of the source and target audios having the same linguistic sentence [16]. *Text-dependent* factor requires the transcribed sentences along with the audio[26].

A classic voice conversion method includes *speech analysis* module, that is, to extract features and analyse the speech signals to create a representational feature which can be modified or transformed for specific uses; *mapping* and *conversion* modules, which compromises of mapping the source and target features for efficient conversion of the speech features and *reconstruction* module focuses on converted features back into a speech signal [21].

Preprint. Under review.

2 Data Corpus

The Common Voice Corpus 14.0 [2] is a vast collection of transcribed speech data for research and development in speech technology. This corpus is designed for Automatic Speech Recognition (ASR) purposes but has broader applications, such as language identification. Common Voice relies on crowd-sourcing for data collection and validation. The most recent release includes data in 29 languages, with 38 languages collecting data in total. Over 88,154 voices are present in the dataset, resulting in 3,279 hours of audio data out of 2484 hours of data that have been validated. It's one of the largest open-access speech corpora for ASR and other audio-related tasks in terms of both data volume and language diversity.

The contributors participate by recording their voices using the Common Voice website or iPhone app. They read sentences displayed on the screen during the recording process. The sentences for recording are typically selected from a pool of text prompts. A validation step is employed to ensure the quality of the recorded data. Other contributors, separate from the original recorders, assess the audio quality by using a simple voting system. They review the audio-transcript pairs and vote to indicate if the recording is correct (up-vote) or incorrect (down-vote).

3 Exploratory Data Analysis

This section will discuss various exploratory data analysis techniques applied to a subset of Common Voice English Corpus 14.0 [2]. We have taken a subset of the dataset in such a way that we will be working on the parallel corpora of the dataset for voice conversion. Exploratory data analysis(EDA) is an approach to studying and analyzing datasets by exploring their various characteristics, discovering patterns, identifying relationships between the features either from the original dataset or generated using feature engineering techniques, and summarizing them through visual tools.

Exploratory Data Analysis (EDA) is crucial for working with audio data due to several reasons. Audio data is information-rich, and EDA helps identify key features that can be extracted for further analysis, such as spectral characteristics, pitch, tempo, and audio fingerprinting. EDA provides a deeper understanding of audio data, revealing domain-specific insights. For instance, EDA can uncover variations in speech patterns, accents, or background noise in speech recognition. Summarizing and visualizing these relevant features aids in selecting the appropriate modelling techniques and algorithms for specific audio data analysis.

3.1 Audio Analysis

This section discusses the various audio analysis techniques to learn more about the audio data.

3.1.1 Average Energy

Average energy, which measures the overall loudness of an audio recording by averaging the squared audio samples, can be used in EDA to identify volume variations, noisy segments, and sudden loudness changes. It can also be used as a feature to distinguish voices based on their loudness and intensity, which can help transform these vocal characteristics during voice conversion.

3.1.2 Average Amplitude

Average amplitude, denoting the mean magnitude of audio signal samples, serves as a metric for assessing the typical sample value magnitude, indicative of an audio recording's usual loudness level. This metric facilitates the analysis of volume variations and distortions during EDA, offering insights into audio loudness patterns. In voice conversion, it aids in preserving or modifying the perceived volume characteristics of the voice, contributing to the attainment of natural-sounding voice transformations.

3.1.3 Average Pitch

Average pitch, the mean fundamental frequency of an audio signal, indicates the speaker's typical pitch or tone. In EDA, Average pitch can be used to analyze the speaker's typical vocal pitch and identify variations in voice characteristics. The average pitch is valuable for capturing the speaker's

fundamental frequency. Preserving or modifying this pitch information is essential to ensure that the converted voice retains a natural and appropriate tone, transforming sound convincing.

3.1.4 Estimated Syllable Count

Estimated syllable count approximates the number of syllables in text or speech, offering a metric for evaluating speech complexity and rhythm. This measure allows for the assessment of speech patterns and styles. It proves valuable in preserving the natural pace and rhythm of speech during voice conversion, enhancing the authenticity and coherence of the converted voice.

3.1.5 Estimated Rate of Speech

The Estimated Rate of Speech quantifies spoken language's pace, denoting syllables articulated per second. This metric aids in tempo discernment and identification of nuances in audio data, like accents and emotional shifts. In voice conversion, it aligns the transformed voice with the original speech's tempo, preserving rhythm and intonation, thereby enhancing authenticity while retaining the speaker's distinct characteristics.

Sentence	it had a diameter of about thirty yards
Age	fourties
Gender	male
Accent	United States English

Average Energy	0.013
Average Amplitude	0.062 in normalised units
Average Pitch	48.776 Hertz
Estimated Syllable Count	8
Rate Of Speech	1.893 syllables/second
Audio Duration	4.224 second

Figure 1: Dataset Feature

3.2 Audio Visualization

3.2.1 Waveform

A waveform is a graphical representation of an audio signal, depicting changes in signal amplitude over time. It provides insights into the signal's structure and attributes, facilitating pattern recognition and the detection of distinctive characteristics. For voice comparison, the waveform is valuable in assessing conversion quality by visually comparing the original and transformed waveforms. This aids in identifying anomalies, phase differences, or artifacts in the converted speech.

3.2.2 Pitch contour

A pitch contour graphically represents pitch variations in an audio signal, reflecting its melodic and intonational qualities. This visual tool aids in analyzing pitch patterns, allowing the identification of prosodic nuances, often indicative of emotional and intentional aspects of speech. Importantly, the pitch contour preserves the original speaker's prosodic attributes during conversion, enhancing the authenticity and emotional expressiveness of the transformed voice while safeguarding the intended content and speaking style.

3.2.3 Spectrogram

A spectrogram is a visual representation that displays how the frequency spectrum of an audio signal changes over time. It reveals variations in energy across frequencies, aiding the identification of patterns, transient sounds, and temporal aspects in audio data. Spectrograms are essential for examining spectral attributes in both the original and desired voices. Aligning these spectrograms enhances voice transformation precision, preserving spectral features, and resulting in a transformed voice with improved naturalness and audio quality.

3.2.4 Pre-Emphasized waveform

A pre-emphasized waveform is an audio signal subject to a filter that amplifies higher frequencies while attenuating lower ones. This technique enhances the intelligibility of high-frequency elements, like consonants and speech characteristics. Employing pre-emphasized waveforms aids in discerning and analyzing critical acoustic details in high-frequency components. This, in turn, improves the preservation and clarity of high-frequency attributes during the conversion process, leading to an overall enhancement in the quality of the transformed voice.

3.2.5 Chromogram

A chromogram is derived from a spectrogram and reveals musical pitch distribution in audio over time. It identifies pitch presence, chord progressions, melody, harmony, and key detection. It aids in analyzing music content, making it useful in tasks like music information retrieval and genre classification.

3.2.6 Mel-Frequency Cepstral Coefficients

Mel-Frequency Cepstral Coefficients (MFCCs) represent the short-term power spectrum of an audio signal, replicating the human auditory system's response to sound. Derived from the Fourier Transform, MFCCs efficiently capture essential spectral features, making them prominent in speech and audio processing. MFCCs provide a concise portrayal of spectral content, facilitating analysis of acoustic characteristics like timbre and phonetic information, with particular significance in audio-related research. They are crucial for analyzing and modifying phonetic information during voice conversion, enabling distinctions between source and target voices while preserving overall speech quality.

3.2.7 Spectral Centroid

Spectral Centroid, a metric, indicates the central location in an audio signal's power spectrum, revealing the frequency range with the most acoustic energy. It's instrumental for understanding audio's spectral attributes and characterizing distinct features among voices. Comparing spectral centroids aids in achieving precise, high-quality voice conversions by highlighting differences in timbre and pitch.

3.2.8 Spectral Bandwidth

Spectral Bandwidth, a metric, quantifies a power spectrum's width in an audio signal, offering insights into frequency component dispersion and the signal's frequency range. It's pivotal for assessing audio's spectral characteristics, distinguishing complexity and purity based on bandwidth. Understanding spectral bandwidth differences between voices aids in achieving precise voice transformation while preserving unique timbre and attributes.

3.3 Text Analysis

3.3.1 Sentence Length Histogram and Maximum and Average Sentence Length

Sentence Length Histogram: It is a graphical representation that displays the frequency or count of sentences at different lengths within a text or document. It provides a visual summary of sentence structure, helping to identify patterns, variations, or anomalies in sentence length distribution.

Maximum and Average Sentence Length: This code calculates and displays the maximum and average sentence lengths in the dataset, giving an idea of the overall text complexity and length.

3.3.2 Most Common Word Frequency and Word Cloud

Most Common Word Frequency: This refers to the frequency with which words appear in a text or document. It involves counting the occurrence of each word and determining which words occur most frequently.

Word Cloud: A word cloud is a visual representation of text data in which words are displayed in varying sizes based on their frequency in the text.

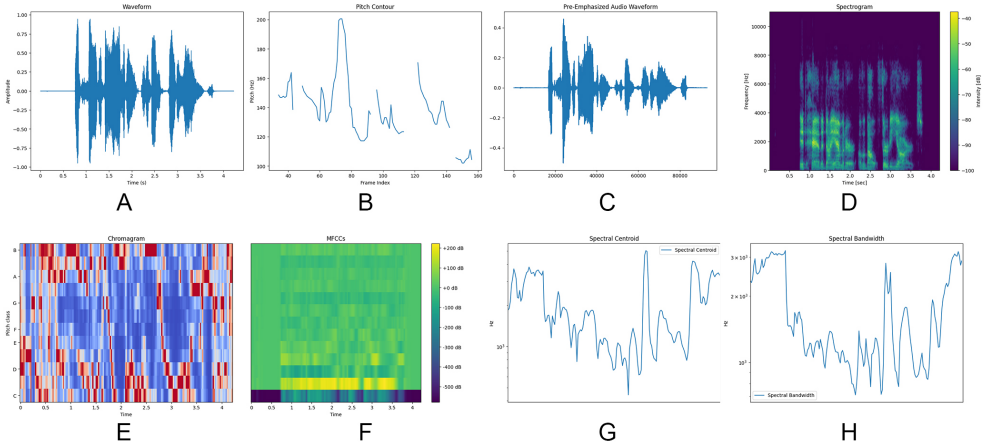

Figure 2: Dataset Visualization A: Waveform, B: Pitch Contour, C: Pre-Emphasized Waveform, D: Spectrogram After Pre-Emphasis, E: Chromagram, F: MFCCs (Mel-Frequency Cepstral Coefficients), G: Spectral Centroid, H: Spectral Bandwidth

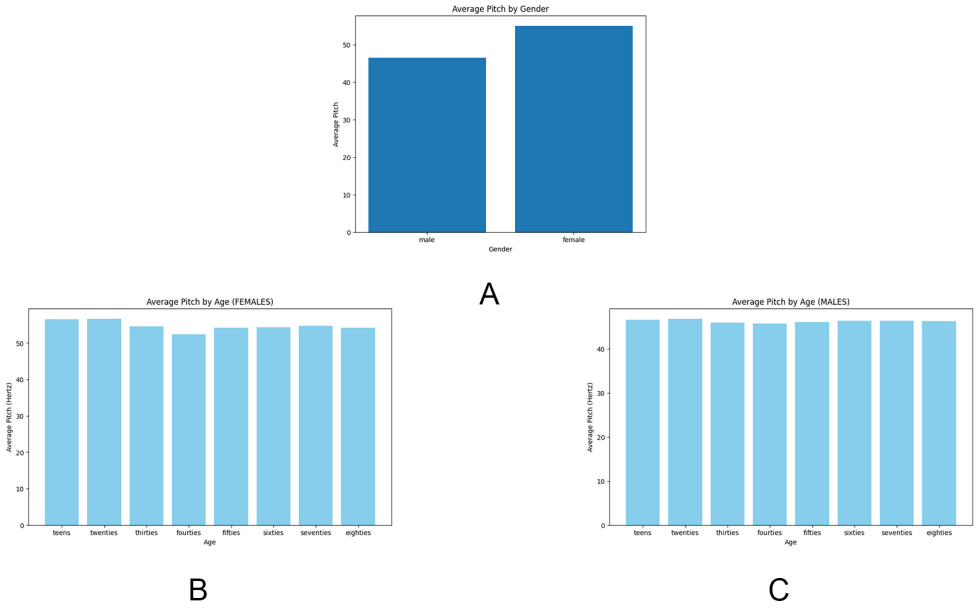

Figure 3: Dataset Visualization A: Average Pitch Comparison between Male and Female B: Pitch Comparison by Age in Females, C: Pitch Comparison by Age in Males.

3.4 Statistics Analysis

3.4.1 Age Distribution

It counts the frequency of each unique value in the 'age' column, creating a bar plot to visualize the distribution of ages in the dataset.

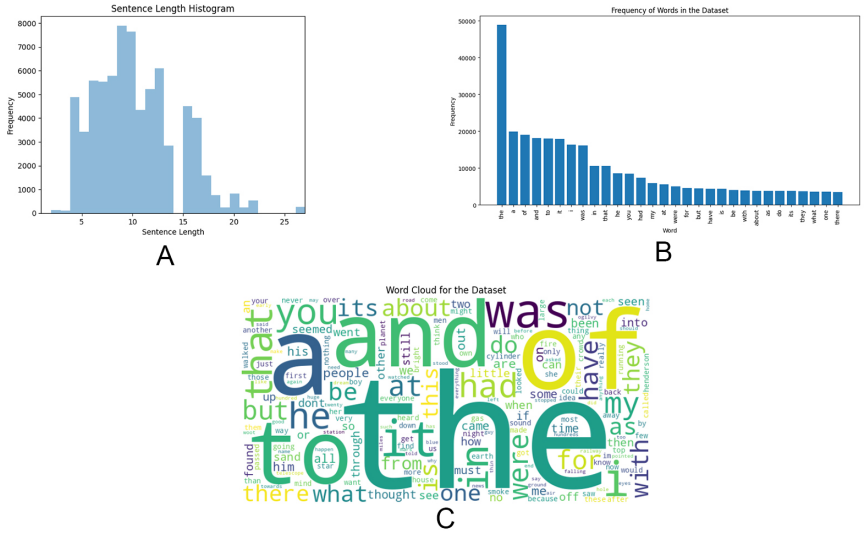

Figure 4: Text Analysis: A: Length of sentences, B: Most Common occurring words, C: Word Cloud

3.4.2 Gender Distribution

It counts the frequency of each unique value in the 'gender' column and presents the distribution of genders through a bar plot.

3.4.3 Accent Distribution

It calculates the frequency of each unique value in the 'accents' column, visualizing the distribution of accents with a bar plot.

3.4.4 Sentence Frequency Plot

This code generates a histogram displaying the frequency distribution of sentences, helping understand the distribution of sentence occurrences in the dataset.

4 Existing Work

This section discusses different approaches and methods of voice conversion mapping, focusing mainly on statistical and classical mapping, and a brief overview of different neural network approaches has also been discussed.

4.1 Codebook mapping

Codebook mapping[14], in the context of data processing and vector quantization, can be improved in several ways. One approach involves storing a difference vector between source and target centroids as a codebook (VQ-DIFF)[13], which can help capture more variability and reduce quantization error. Additionally, soft clustering techniques can be applied, such as joint-density (JD) modelling, to associate source and target codebook vectors. This involves stacking source and target vectors and then estimating joint codebook vectors using clustering algorithms. Notably, JDVQ-DIFF can generate samples that were not present in the target training data, which JDVQ cannot accomplish. However, it's essential to note that JDVQ tends to exhibit high quantization error, and both JDVQ and JDVQ-DIFF are prone to generating discontinuous feature sequences. These methods offer ways to optimize codebook mapping and improve its performance in various applications.

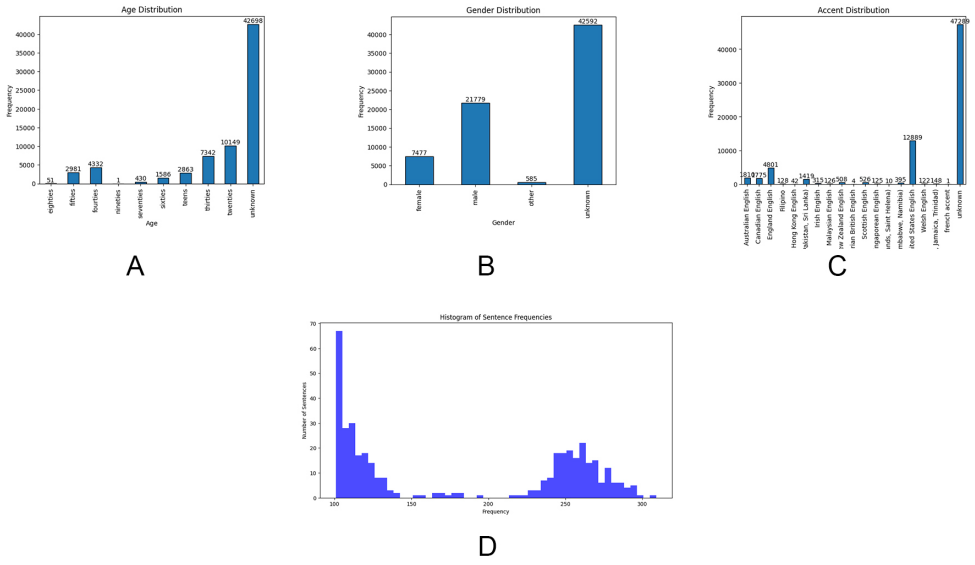

Figure 5: Statistics Analysis: A: Age Distribution, B: Gender Distribution, C: Accent Distribution, D: Histogram Of Sentence Frequencies

4.2 Frequency warping mappings

Various methods have been explored for voice conversion. Initially, Valbret et.al. [33] proposed frequency warping using pre-computed warping functions between source and target speakers, adjusting spectral tilt. Some studies directly model and manipulate formant frequencies and bandwidths to match target formants, while others cluster the acoustic space into different classes, applying non-linear frequency warping. Sündermann et.al. [24] investigated vocal tract length normalization (VTLN) approaches. Erroetal[7] extended VTLN to multiple classes. Přibilová and Přibil [19] experimented with various linear and non-linear warping functions. Erro and Moreno [6] introduced weighted frequency warping (WFW), enhancing flexibility. Toda et. al. [30] proposed converting the source spectrum using a GMM and then warping it to preserve spectral details.

In addition to formant frequencies, the average energy of spectral bands is vital for speaker individuality. To address this, an energy correction filter was introduced, and amplitude scaling with a shift value was proposed. Some frequency warping functions can be reformulated as weighted linear mapping approaches. However, these approaches often involve constraints to prevent over-fitting, which may limit their ability to mimic very different voices.

Numerous extensions of the frequency warping approach have been developed, including combinations with GMMs, dictionary-based methods, and maximizing spectral correlation, demonstrating the versatility of voice conversion techniques.

4.3 Dictionary mapping

To address discontinuities in generated features, a unit-selection (US) paradigm is employed. US methods utilize a target cost and a concatenation cost, ensuring that neighboring target features are most similar to each other, often referred to as frame-selection (FS). The goal is to minimize a cost function, balancing fitting accuracy and spectral continuity [20] [31] [12]. Dynamic programming, like Viterbi, is used to find the optimal target sequence, which can be used for aligning frames or as a mapping function.

US/FS techniques can be adapted for text-independent, non-parallel voice conversion (VC) systems[25] [27]. In this variation, a vector is compared to a target training vector in the dictionary to compute the target cost. These techniques can also be used to parallelize training data.

Various combinations and variants of US/FS are proposed with other mapping approaches like dictionary mapping [8], GMM mapping [4], k-histogram [32], exemplar-based VC [34], and grid-based approximation [3]. However, these methods may have limitations and can produce discontinuous features, particularly when working with a small number of training utterances, which is common in VC tasks.

4.4 Neural network mapping

Artificial Neural Networks (ANNs) are powerful tools used in various applications, including voice conversion (VC). They require careful training and design to avoid issues like local minima. ANNs, like Gaussian Mixture Models (GMMs), can approximate various functions[28]. GMMs achieve non-linearity differently through the sum of class-based linear transformations, while ANNs use non-linear activation functions[11]. Different ANN architectures, such as recurrent architectures[17] [23], are applied in VC.

To improve training, techniques like pre-training with initial parameter values[5], rectified linear units[10], and dropout[22] have been used. These methods address challenges like vanishing gradients and local minima[9], which are crucial for VC tasks with limited data. Additionally, a sequence error minimization[35] approach using Recurrent Neural Networks (RNNs) captures the temporal context in VC applications.

5 Dataset and Feature Extraction

5.1 Audio Data

A dataset comprising 108 audio samples, with 54 from each of the two speakers, has been curated for this study. Within this dataset, one speaker is designated as the source, while the other assumes the role of the target. Both speakers articulate identical sets of 54 English sentences. This meticulously designed dataset serves the dual purpose of training and evaluating the effectiveness of the generated models. By employing consistent sentences and parallel utterances, the dataset provides a robust foundation for exploring the intricate dynamics of voice conversion, enabling a comprehensive assessment of the model's performance across various linguistic contexts and acoustic scenarios.

5.2 Feature Extraction and Alignment

For a voice conversion pipeline, it needs to modify speaker-specific acoustic features contained in the speech signal. In this paper, we focused on essential sound features like spectrogram, Mel-frequency Cepstral Coefficients (MFCC), and Mel-spectrogram. We chose these features because they do not consider the speech model and are independent of the linguistic context. In the initial phase, accurately aligning the timing of sound recordings from both the source and target speakers was crucial. This alignment ensures that the features in the sound recordings match up well, making the training process smooth and effective. Dynamic time warping (DTW) has been used to derive the time-aligned features from the parallel utterances of the source and target speakers. Ensuring precise timing helps the model understand and absorb subtle variations in vocal characteristics, ultimately improving the overall performance of the voice conversion system.

5.3 Reconstruction from extracted features

5.3.1 Spectogram

The Griffin-Lim algorithm takes a magnitude (or power, or log-power) spectrogram as input to reproduce an audio signal consistent with the desired spectrogram. The Griffin-Lim algorithm attempts to estimate the missing or distorted phase information and reconstruct the signal by iteratively refining the time-domain signal using the estimated phase. The basic algorithm is an iteration that takes the prior estimate of the complex spectrogram, applies the STFT backward and forward, and reapplies the desired magnitude. In each back-and-forward step, information leaks between time-frequency bins taking the estimate closer to a consistent estimate. ISTFT(Inverse short-time Fourier transform) can also be used to reconstruct audio signals from its spectrograms

5.3.2 Mel-Spectogram

The mel spectrogram is a representation of the power spectrum of a signal, typically used in audio processing tasks. Reconstruction involves using the inverse mel filter bank to convert the mel spectrogram back into a linear-scale spectrogram and then using the Griffin-Lim algorithm to reconstruct the time-domain audio signal from its frequency-domain representation.

5.3.3 MFCC's

MFCC's (Mel-Frequency Cepstral Coefficients) represent the short-term power spectrum of a sound signal. Various processes like Pre-Emphasis, Framing, Windowing, FFT(Fast Fourier Transform), and DCT(Discrete Cosine Transform) are used to compute these coefficients. Converting MFCCs back to audio involves inverse MFCC transformation. It should be noted that while computing MFCC, many other features of the signal are ignored and thus some information loss occurs during the original MFCC extraction process. Thus, the reconstruction might not be perfect due to the non-invertibility of the MFCC transformation.

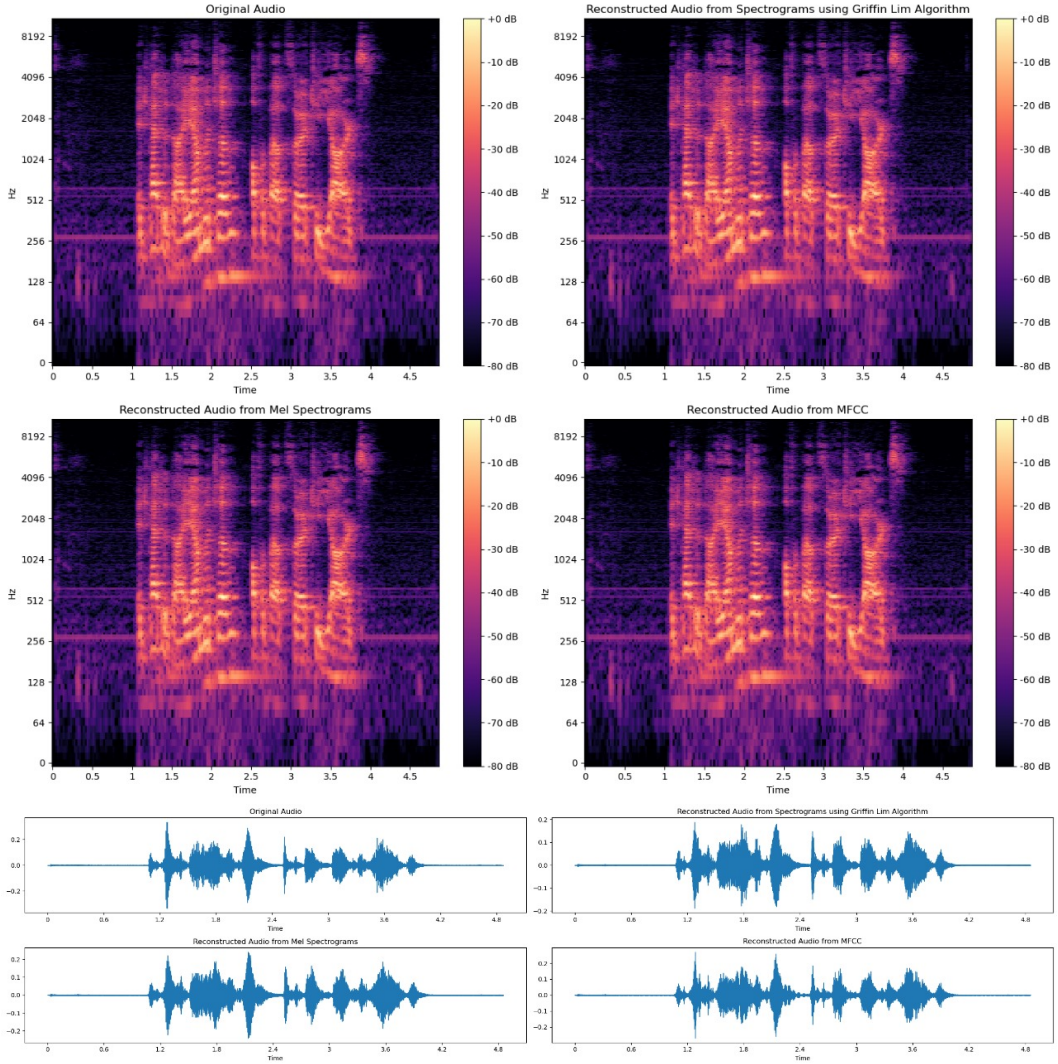

Figure 6: Differences in Spectrogram and Wave plot when an audio is reconstructed from different features

6 Methodologies Used

This section delineates the diverse methodologies and procedural steps employed in our endeavors toward voice conversion. These methodologies encompass the creation of a features matrix, wherein different techniques such as dictionary mapping, non-negative matrix factorization, and Gaussian Mixture Model (GMM) application utilizing Maximum Likelihood Parameter Generation (MLPG) were systematically applied.

6.1 Speech analysis

In relation to the dataset mentioned earlier, we calculated the Mel-frequency cepstral coefficients (MFCC) for individual elements. The MFCC was computed with a designated order of 13, utilizing segmentation intervals spanning from 22 to 33 milliseconds. The choice of order and segmentation parameters in MFCC computation involves a trade-off between capturing fine details and managing computational complexity. Furthermore, zero-padding was employed, and temporal alignment was executed concerning both the source and target. Consequently, this process culminated in the formation of a cohesive feature matrix.

6.2 Mapping function

6.2.1 Direct Mapping and Linear Regression

This method involves creating a mapping between aligned source and target feature vectors and applying this mapping to convert source features to corresponding target features. The code processes source and target audio by extracting both spectrogram and Mel-frequency cepstral coefficients (MFCC) features, aligns these features, and generates separate mappings for spectrogram and MFCC features. Finally, the mappings are utilized to convert aligned source features to target features for both types of extracted features.

In linear regression model, features from the source speaker and the corresponding features from the target speaker are used to train a model capable of transforming new source features into their corresponding target features. It is important to note that the dimensions of the input features during training and conversion must be consistent for the model to produce meaningful results.

The resulting feature matrix is used to construct the resulting audio using the above-mentioned methods.

6.2.2 Non-Negative matrix factorization

In this method, we combined Non-Negative Matrix Factorization (NMF), Dynamic Time Warping (DTW), and the Griffin-Lim algorithm. The process involves loading source and target audio files, extracting Mel-frequency cepstral coefficients (MFCC) features, aligning features temporally through DTW, learning dictionaries using NMF, and converting a new source audio using these dictionaries. The resulting converted audio is then saved as a new file. This approach enables the transformation of voice characteristics from a source speaker to those of a target speaker, offering a method for voice modification and adaptation.

6.2.3 GMM using Maximum Likelihood Parameter Generation (MLPG)

A Gaussian mixture model(GMM) of the joint probability density of source and target features is employed as a Mapping function for performing spectral conversion between speakers. The conventional method includes converting frame-by-frame spectral parameters based on MSE. Although reasonably effective, deterioration of speech quality is caused by 2 problems i.e. the time-independent mapping and the over-smoothing effect.

The time-independent mapping means that the model is considering only static features implying that it is only based on a global pattern across the trajectories. However appropriate spectral movements are not always caused by the frame-based conversion process. Although two trajectories seem similar, they sometimes have different local patterns. Such differences are often observed because the correlation of the target feature vectors between frames is ignored in conventional mapping. In order to realize appropriate spectral movements, we consider the feature correlation between frames

by applying a parameter generation algorithm with dynamic features just like in HMM(Hidden Markov Models). This idea makes it possible to estimate an appropriate spectrum sequence in view of not only static but also dynamic characteristics. Thus, we introduce these dynamic features by MLE(Maximum likelihood estimation) of Spectral parameter trajectory. The likelihood function is introduced by using 2D source and target feature vectors for each time frame consisting of 1D dimensional static and dynamic features at the particular frame.

The converted spectra are excessively smoothed by statistical modeling. Statistical modeling often removes the details of spectral structures. This smoothing undoubtedly causes error reduction of the spectral conversion. However, it also causes the quality degradation of the converted speech because the removed structures are still necessary for synthesizing high-quality speech. This over-smoothing effect can be alleviated by considering the Global variance feature of the converted spectra. we calculate GV utterance by utterance. Then we apply the same method of MLE on the likelihood functions defined on two probablity density functions for the target static and dynamic feature vectors and for the GV of the target static feature vectors. Using more mixture components for modeling the probability density also alleviates the over-smoothing effect but it may cause overfitting.

Thus, This technique enhances the quality of recovered signal while mapping source to target effectively.[29]

6.3 Reconstruction

6.3.1 WORLD vocoder

The WORLD vocoder is a speech analysis and synthesis system for high-quality voice and signal processing. It uses a spectral modelling approach to represent and reconstruct speech signals, including pitch, spectral envelope, and aperiodicity parameters. The synthesis process involves converting these parameters back to the time domain using the overlap-add method, resulting in high-quality reconstructed audio focusing on preserving the original speech's natural characteristics.[15]

6.3.2 Maximum Likelihood Parameter Generation and Spectrum Distortion Function (MLSADF)

MLSADF operates as a synthesizer based on formants, utilizing a filter to shape the spectral envelope of the speech signal. In this framework, the spectral envelope is represented by a set of coefficients that describe the filter's configuration. After the spectral envelope estimation is completed, the synthesis of the speech signal is carried out by filtering a source excitation through the spectral envelope filter. The MLSADF synthesizer excels in producing high-quality speech signals distinguished by a natural and expressive prosody.[18]

6.4 Evaluation

6.4.1 Mel Cepstral Distortion

Mel Cepstral Distortion (MCD) is commonly used in speech and audio processing to quantify the difference between two sets of Mel-frequency cepstral coefficients (MFCCs). It is a metric that reflects the dissimilarity between the spectral characteristics of two audio signals. MCD is particularly relevant in speaker verification, speech recognition, and audio quality assessment scenarios. Given below, are MCD results for 3 source-target pairs.

Source - Target Pair	MFCC Order	GMM Components	Number of Audio	Mel Cepstral Distortion
C1-C2	30	32	54	47.212
C3-C4	30	32	67	43.712
C5-C6	30	32	110	52.408

Figure 7: Mel Cepstral Distortion results for 3 source-target pairs

6.4.2 Pixel-wise reference of Spectrogram

A spectrogram is a visual representation of the spectrum of frequencies of a signal as they vary with time.

Given below, there are three Spectrograms of source, target and converted audio files.

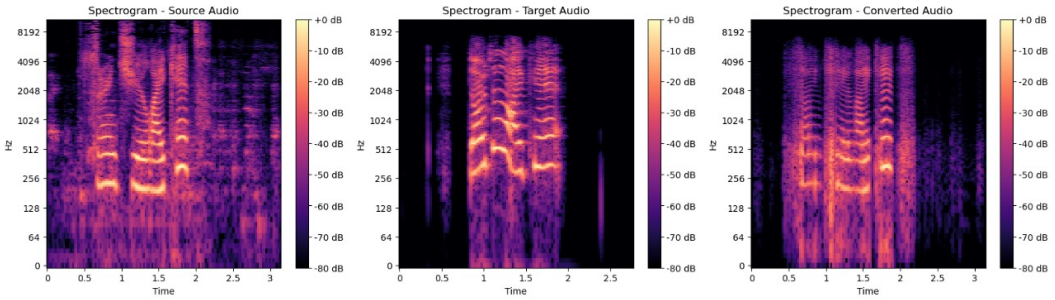

Figure 8: Pixel-Wise Reference (Spectrogram) : A: Source Audio, B: Target Audio, C: Converted Audio

7 Future Plans

This section will discuss the future implementations we may aim to achieve for this project. These implementations are one of the classical machine learning techniques, which has their own drawbacks which cannot be overcome by different types of feature extraction processes. The voice conversion process can be improved by using multiple modern techniques, such as Deep Learning and GANs, which can model the data perfectly and extract features more efficiently than the hand-crafted features used in Audio-Signal preprocessing techniques. The Deep Learning models include RNN, LSTM and Transformers, and GANs include STARGANs and CYCLEGANs. These can be employed on this dataset to achieve a superior performance.

References

[1] Rosana Ardila et al. "Common Voice: A Massively-Multilingual Speech Corpus". In: *CoRR* abs/1912.06670 (2019). arXiv: 1912.06670. URL: http://arxiv.org/abs/1912.06670.

[2] Rosana Ardila et al. *Common Voice: A Massively-Multilingual Speech Corpus*. 2020. arXiv: 1912.06670 [cs.CL].

[3] Hadas Benisty, David Malah, and Koby Crammer. "Sequential voice conversion using grid-based approximation". In: *2014 IEEE 28th Convention of Electrical & Electronics Engineers in Israel (IEEEI)*. IEEE. 2014, pp. 1–5.

[4] Helenca Duxans et al. "Including dynamic and phonetic information in voice conversion systems". In: *Proc. of the ICSLP'04*. 2004.

[5] Dumitru Erhan et al. "Why does unsupervised pre-training help deep learning?" In: *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings. 2010, pp. 201–208.

[6] Daniel Erro and Asunción Moreno. "Weighted frequency warping for voice conversion." In: *Interspeech*. 2007, pp. 1965–1968.

[7] Daniel Erro, Eva Navas, and Inma Hernáez. "Iterative MMSE estimation of vocal tract length normalization factors for voice transformation". In: *Thirteenth Annual Conference of the International Speech Communication Association*. 2012.

[8] Kei Fujii, Jun Okawa, and Kaori Suigetsu. "High-individuality voice conversion based on concatenative speech synthesis". In: *International Journal of Electrical and Computer Engineering* 1.11 (2007), pp. 1625–1630.

[9] Xavier Glorot and Yoshua Bengio. "Understanding the difficulty of training deep feedforward neural networks". In: *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings. 2010, pp. 249–256.

[10] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. "Deep sparse rectifier neural networks". In: *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings. 2011, pp. 315–323.

[11] Rabul Hussain Laskar et al. "Comparing ANN and GMM in a voice conversion framework". In: *Applied Soft Computing* 12.11 (2012), pp. 3332–3342.

[12] Ki-Seung Lee. "A unit selection approach for voice transformation". In: *Speech Communication* 60 (2014), pp. 30–43.

[13] Hiroshi Matsumoto and Yasuki Yamashita. "Unsupervised speaker adaptation from short utterances based on a minimized fuzzy objective function". In: *Journal of the Acoustical Society of Japan (E)* 14.5 (1993), pp. 353–361. DOI: 10.1250/ast.14.353.

[14] Seyed Hamidreza Mohammadi and Alexander Kain. "An overview of voice conversion systems". In: *Speech Communication* 88 (2017), pp. 65–82. ISSN: 0167-6393. DOI: https://doi.org/10.1016/j.specom.2017.01.008. URL: https://www.sciencedirect.com/science/article/pii/S0167639315300698.

[15] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa. "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications". In: *IEICE TRANSACTIONS on Information and Systems* 99.7 (2016), pp. 1877–1884.

[16] A. Mouchtaris, J. Van der Spiegel, and P. Mueller. "Non-parallel training for voice conversion by maximum likelihood constrained adaptation". In: *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 1. 2004, pp. I–1. DOI: 10.1109/ICASSP.2004.1325907.

[17] Toru Nakashika, Tetsuya Takiguchi, and Yasuo Ariki. "Voice conversion using RNN pre-trained by recurrent temporal restricted Boltzmann machines". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23.3 (2014), pp. 580–587.

[18] Sanghamitra Nath et al. "VoiCon: a Matlab GUI-based tool for voice conversion applications". In: *International Journal of Computer Applications in Technology* 61.3 (2019), pp. 207–219.

[19] Anna Přibilová and Jiří Přibil. "Non-linear frequency scale mapping for voice conversion in text-to-speech system with cepstral description". In: *Speech Communication* 48.12 (2006), pp. 1691–1703.

[20] Özgül Salor and Mübeccel Demirekler. "Dynamic programming approach to voice transformation". In: *Speech communication* 48.10 (2006), pp. 1262–1272.

[21] Berrak Sisman et al. "An Overview of Voice Conversion and Its Challenges: From Statistical Modeling to Deep Learning". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), pp. 132–157. DOI: 10.1109/TASLP.2020.3038524.

[22] Nitish Srivastava et al. "Dropout: a simple way to prevent neural networks from overfitting". In: *The journal of machine learning research* 15.1 (2014), pp. 1929–1958.

[23] Lifa Sun et al. "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks". In: *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 2015, pp. 4869–4873.

[24] David Sundermann, Hermann Ney, and H Hoge. "VTLN-based cross-language voice conversion". In: *2003 IEEE workshop on automatic speech recognition and understanding (IEEE Cat. No. 03EX721)*. IEEE. 2003, pp. 676–681.

[25] David Sündermann. "Text-independent voice conversion". PhD thesis. München, Univ. der Bundeswehr, Diss., 2008, 2008.

[26] David Sündermann et al. "A first step towards text-independent voice conversion". In: *Proc. of the ICSLP'04*. 2004.

[27] David Sündermann et al. "Text-independent cross-language voice conversion." In: *INTERSPEECH*. 2006.

[28] D. M. (David Michael) Titterington, Adrian F. M. Smith, and U. E. Makov. *Statistical analysis of finite mixture distributions*. Wiley series in probability and mathematical statistics. Wiley, 1985. URL: https://cir.nii.ac.jp/crid/1130282271077501312.

[29] Tomoki Toda, Alan W. Black, and Keiichi Tokuda. "Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory". In: *IEEE Transactions on Audio, Speech, and Language Processing* 15.8 (2007), pp. 2222–2235. DOI: 10.1109/TASL.2007.907344.

[30] Tomoki Toda, Hiroshi Saruwatari, and Kiyohiro Shikano. "Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum". In: *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*. Vol. 2. IEEE. 2001, pp. 841–844.

[31] Alejandro Uriz et al. "Voice Conversion Using Frame Selection". In: *Reporte Interno Laboratorio de Comunicaciones-UNMdP* (2008).

[32] Alejandro José Uriz et al. "Voice conversion using k-histograms and frame selection". In: *Tenth Annual Conference of the International Speech Communication Association*. 2009.

[33] Hélene Valbret, Eric Moulines, and Jean-Pierre Tubach. “Voice transformation using PSOLA technique”. In: *Speech communication* 11.2-3 (1992), pp. 175–187.

[34] Zhizheng Wu et al. “Exemplar-based unit selection for voice conversion utilizing temporal information.” In: *INTERSPEECH*. Lyon. 2013, pp. 3057–3061.

[35] Feng-Long Xie et al. “Sequence error (SE) minimization training of neural network for voice conversion”. In: *Fifteenth Annual Conference of the International Speech Communication Association*. 2014.