Voice Conversion using Classical Machine Learning Techniques

Aindrila Majumder
MT23109

Jafreen Rizvi
MT23040

Kanishk Goel
2021325

Modi Akshay
2021166

Aalokik Singh
2021223

Abstract

This report explores the application of classical machine learning techniques, specifically Gaussian Mixture Models (GMMs), for voice conversion using Mozilla's Common Voice English Corpus V14.0. Voice conversion is the process of transforming a source speaker's voice to match that of a target speaker while preserving linguistic content. We draw inspiration from previous research on GMM-based voice conversion and propose a comprehensive methodology for voice transformation. Our approach involves in-depth spectral feature extraction, advanced maximum-likelihood estimation, and a dynamic adjustment of spectral parameter trajectories. We rigorously evaluate the converted voices for quality and naturalness using objective metrics and subjective listening tests. This project aims to demonstrate the practicality and effectiveness of GMM-based voice conversion techniques and explore new possibilities in the realm of voice processing using classical machine learning methods.

1 Introduction

Voice conversion is a fascinating field in speech processing that involves transforming the characteristics of a source speaker's voice to match those of a target speaker, while preserving the linguistic content. This technology has numerous applications, ranging from personalized text-to-speech systems to privacy in telecommunications. A typical VC application is speaker conversion, in which the voice of a certain speaker (source speaker) is converted to sound like that of another speaker (target speaker).

Recent advancements in voice conversion techniques have seen a shift towards using Gaussian Mixture Models (GMMs) due to their robustness and efficacy in modeling complex voice attributes. The research paper "Voice Conversion using Gaussian Mixture Models" by Alan W Black and others at Carnegie Mellon University presents a comprehensive study on this approach. The paper explores the use of GMMs to map the acoustic space of a source speaker to that of a target speaker, thereby achieving voice conversion.

In our project, we draw inspiration from this research and aim to investigate the potential of classical machine learning techniques, with a particular focus on GMMs, for voice conversion using the English Corpus V14.0 from Mozilla's Common Voice initiative. This initiative, while popular in deep learning contexts, is equally relevant for classical approaches, especially given their advantages in terms of interpretability and computational efficiency.

Our goal is to explore the practicality and effectiveness of GMM-based voice conversion techniques. We propose to evaluate our methods using objective metrics such as spectrogram comparisons and to compare them against the benchmarks established in the referenced study. This approach allows us to

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

not only validate the effectiveness of classical methods but also to explore new horizons in the realm of voice conversion.

2 Exploratory Data Analysis (EDA)

2.1 Dataset Overview

The dataset employed in our study is the English Corpus V14.0 from Mozilla's Common Voice initiative. This dataset is particularly suited for voice conversion tasks, offering a rich collection of audio recordings from diverse speakers, each accompanied by metadata that provides valuable context for analysis.

2.2 Spectrogram Analysis

Spectrograms play a crucial role in voice conversion as they visually represent the spectrum of frequencies of a sound signal as it varies with time. This time-frequency representation is essential for understanding the characteristics of different voices.

- **Visualizing Frequency and Time:** Spectrograms provide a visual representation of the frequency spectrum over time, highlighting how different phonetic components and intonations manifest in the audio signal.
- **Comparative Analysis:** By comparing the spectrograms of different audio samples, we can observe distinct patterns that characterize individual speakers, aiding in the voice conversion process.

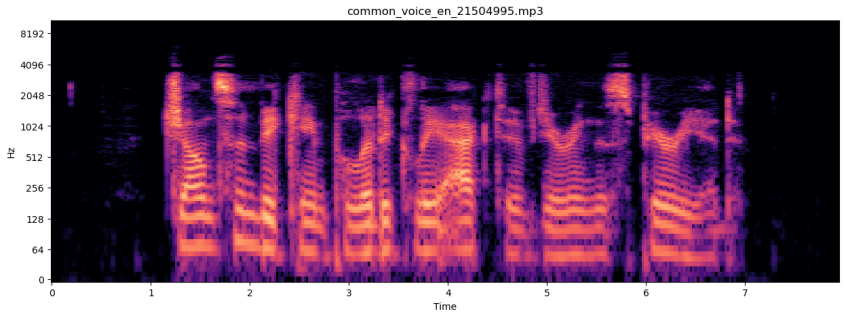

Figure 1: Spectrogram representation of an audio sample.

Figure 1 illustrates a typical spectrogram of an audio sample from the dataset, showcasing the distribution of energy across various frequencies over time.

2.3 Analysis of Mel-Frequency Cepstral Coefficients (MFCCs)

Mel-frequency cepstral coefficients (MFCCs) are a representation of the short-term power spectrum of a sound. MFCC coefficients are derived by applying a series of operations such as FFT, applying mel filters, log and Discrete Cosine Transform (DCT). They capture information about the spectral envelope, which includes details about the shape and distribution of energy in different frequency bands. MFCCs effectively capture the timbral aspects of sound, making them ideal for distinguishing between different voices. MFCCs provide a compact and informative representation of speech that can be changed and transformed to achieve voice conversion.

2.3.1 Zero-Crossing Rate

The zero-crossing rate can be utilized as a basic pitch detection algorithm for monophonic tonal signals. It represents the rate at which the signal changes its sign. In audio, this often corresponds to the rate at which the waveform crosses the zero amplitude line.Voice activity detection (VAD), which determines whether or not human speech is present in an audio segment, also makes use of zero-crossing rates.

2.3.2 Chroma Features

Chroma features are an interesting and powerful representation for music audio in which the entire spectrum is projected onto 12 bins representing the 12 distinct semitones (or chroma) of the musical octave.Chroma features are widely used in audio signal processing and music analysis. They represent the distribution of energy in different pitch classes (i.e., musical notes) over time.

3 Libraries Used and Code Explanation

In our voice conversion project, we utilize a range of Python libraries, each serving a specific purpose in the realm of machine learning and voice processing. These libraries facilitate various stages of our project, from data handling and preprocessing to feature extraction, voice conversion modeling, and audio synthesis.

3.1 Detailed Overview of Libraries

- **nnmnkwii:** Offers tools for speech synthesis and voice conversion. It is particularly effective for handling spectral and prosodic features, crucial for manipulating voice characteristics.
- **pyroomacoustics:** Used for audio signal processing, focusing on acoustics simulation and implementing audio processing algorithms, essential for creating realistic audio outputs.
- **pyworld:** A library for speech signal processing, crucial for decomposing speech signals into fundamental frequency, spectral envelope, and aperiodicity, key components in reconstructing speech sounds.
- **pysptk:** The Python Speech Toolkit provides various speech signal processing tools, including robust feature extraction capabilities like MFCCs extraction.
- **Librosa:** A Python package for music and audio analysis, Librosa provides the tools for audio and music analysis, including feature extraction, such as MFCCs, and signal processing.
- **NumPy:** Fundamental package for scientific computing in Python, NumPy offers support for large, multi-dimensional arrays and matrices, along with a collection of mathematical functions to operate on these arrays.
- **SciPy:** An ecosystem of open-source software for mathematics, science, and engineering, SciPy is used for technical and scientific computing tasks like signal processing.
- **Scikit-learn:** A simple and efficient tool for data mining and data analysis, Scikit-learn is used for implementing machine learning algorithms, particularly for tasks like classification and clustering.
- **TensorFlow/Keras:** TensorFlow, along with its high-level API Keras, is used for building and training deep learning models. It offers flexible and comprehensive tools for complex voice conversion models using neural networks.
- **Matplotlib:** A plotting library for Python, Matplotlib is used for creating static, interactive, and animated visualizations in Python, particularly useful for visualizing features like spectrograms and MFCCs.

4 Failed Procedures

4.1 Voice Feature Conversion using Regression

One approach we attempted for voice conversion involved utilizing Dynamic Time Warping (DTW) and regression techniques. Here is an overview of the procedure:

5 Failed Procedures

5.1 Voice Feature Conversion using Regression

One approach we attempted for voice conversion involved utilizing Dynamic Time Warping (DTW) and regression techniques. Here is an overview of the procedure:

1. We calculated the DTW distance between each vector in the source feature set (source_features) and all vectors in the target feature set (target_features). DTW measures the similarity between two sequences that may vary in time or speed.

2. The code aimed to find the closest match in the target_features for each vector in source_features by selecting the target feature vector with the minimum DTW distance.

3. We employed Ridge Regression (linear regression with L2 regularization) to establish a mapping between the source and the best-matching target feature vectors. The regression was performed by fitting a model (regression_model) using the source feature and its corresponding best-matching target feature.

4. The procedure attempted to predict the converted feature vector for each source feature vector using the trained regression model and stored these converted feature vectors in the converted_features list.

5. Ultimately, the function was intended to return an array (converted_features) containing the converted feature vectors of the source voice.

Despite these efforts, we encountered significant issues with this approach. The updated audio generated using this method exhibited gibberish voice and a substantial amount of noise, rendering it unusable for our purposes.

5.2 GMM Adaptation for Voice Conversion

Another technique we explored was GMM adaptation for voice conversion and similar applications. This approach aimed to adapt a source Gaussian Mixture Model (GMM) using information from target features. The adaptation process involved updating the means of the GMM based on the target data and a defined adaptation factor. The goal was to align the statistical parameters of the source GMM with characteristics observed in the target data, facilitating better modeling or transformation of source features to match those of the target.

Unfortunately, this approach also yielded unsatisfactory results. The updated audio produced using the GMM adaptation technique exhibited gibberish voice and excessive noise, making it impractical for our voice conversion objectives.

These failed procedures underscore the complexity of voice conversion and the challenges in achieving high-quality results. Further research and experimentation are needed to develop more effective methods for this task.

Despite these efforts, we encountered significant issues with this approach. The updated audio generated using this method exhibited gibberish voice and a substantial amount of noise, rendering it unusable for our purposes.

5.3 GMM Adaptation for Voice Conversion

Another technique we explored was GMM adaptation for voice conversion and similar applications. This approach aimed to adapt a source Gaussian Mixture Model (GMM) using information from target features. The adaptation process involved updating the means of the GMM based on the target data and a defined adaptation factor. The goal was to align the statistical parameters of the source GMM with characteristics observed in the target data, facilitating better modeling or transformation of source features to match those of the target.

Unfortunately, this approach also yielded unsatisfactory results. The updated audio produced using the GMM adaptation technique exhibited gibberish voice and excessive noise, making it impractical for our voice conversion objectives.

These failed procedures underscore the complexity of voice conversion and the challenges in achieving high-quality results. Further research and experimentation are needed to develop more effective methods for this task.

6 Methodology

Our methodology for voice conversion is grounded in advanced statistical modeling techniques, focusing on the precise transformation of spectral characteristics from a source speaker to a target speaker. This process involves several sophisticated steps, each contributing to the overall efficacy and quality of the voice conversion.

6.1 Gaussian Mixture Model (GMM)

We have employed a Gaussian Mixture Model (GMM) to preserve the non-linear relationship between audio features. Since audio features are not linear in nature, linear methods like Support Vector Machines (SVM) and simple mapping functions may not yield optimal results. GMM models are capable of preserving non-linear relationships globally while still modeling linear relationships locally within a cluster. We extracted joint features from the source and target speakers and trained these features using the GMM. The GMM allows us to model the probability distribution of a random variable, enabling us to estimate model parameters (α, μ, and Σ) using the Expectation Maximization (EM) algorithm.

In the testing process, we utilize regression to predict feature Y for a given input feature set X from the source speaker using the trained GMM model. Approximately 50 parallel samples of source and target audios were used to train the GMM model.

6.2 Speech Analysis

The initial step in our GMM-based voice conversion model involves the analysis of speech signals from both the source and target speakers. The speech signals are divided into frames, and we compute delta features from Mel-Frequency Cepstral Coefficients (MFCCs), as well as features such as Zero-Crossing Rate (ZCR) and chroma. To ensure consistency in signal length, longer signals are trimmed.

6.3 Mapping

For spectral conversion between speakers, we employ a Gaussian Mixture Model (GMM) representing the combined probability density of source and target characteristics. Instead of frame-by-frame conversion using minimal mean square error, we propose a conversion technique based on the maximum-likelihood estimate (MLE) of a spectral parameter trajectory. MLE seeks to estimate parameters that best describe the transformation of spectral characteristics from a source to a target. It aims to find the most probable sequence of spectral parameters that could have generated the converted spectrum, given the source and target spectra.

6.4 Reconstruction

For the reconstruction process we have used synthesis function in pysptk library.After we get the converted signal . MLSADF in pystk take converted features and generate MLSADF coefficients and synthesis function produce a new synthesized audio using those coefficients and input audio signal .

6.5 1.Speech Analysis :

- **In-depth Feature Extraction:** The first step involves an extensive extraction of spectral features from the audio samples of both source and target speakers. We focus on extracting Mel-Frequency Cepstral Coefficients (MFCCs), spectral envelopes, and other relevant features that define the unique voice characteristics of each speaker.
- **Building the GMM-Based Model:** Gaussian Mixture Models (GMMs) are employed to model the joint probability distribution of the source and target spectral features. This modeling is crucial for capturing the nuanced relationship between different voice characteristics, allowing for a more natural and effective voice conversion.

6.6 Advanced Maximum-Likelihood Estimation

- **Precise Parameter Estimation:** Using maximum-likelihood estimation, we meticulously determine the parameters of the GMM. This statistical approach ensures that our model accurately reflects the spectral characteristics of both source and target speakers, providing a solid foundation for the conversion process.
- **Focus on Spectral Parameter Trajectory:** A key aspect of our methodology is the emphasis on the spectral parameter trajectory. This involves analyzing and modeling how spectral features evolve over time in speech. By capturing these dynamic changes, our voice conversion process is not only able to transform static voice characteristics but also replicate the natural temporal variations in speech, leading to more lifelike and coherent voice conversions.

6.7 Comprehensive Voice Conversion Process

1. **Transformation Application:** The conversion process involves applying the GMM-based models to the extracted spectral features of the source speaker. This step transforms these features to align closely with those of the target speaker, effectively altering the voice characteristics while preserving the linguistic content.
2. **Dynamic Adjustment:** Special attention is given to maintaining the naturalness and dynamism in speech. Adjustments are made to ensure that the converted speech not only matches the target voice in terms of spectral qualities but also in how these qualities change during natural speech.

6.8 Rigorous Evaluation and Refinement

- **Quality Assessment:** The converted voices undergo rigorous evaluation, where their quality and naturalness are assessed through both objective metrics and subjective listening tests.
- **Iterative Refinement:** Based on the feedback and evaluations, we engage in an iterative process of refinement, fine-tuning the model parameters and conversion techniques to achieve the highest possible quality in our voice conversion outputs.

7 Results

7.1 Qualitative Analysis

In our qualitative analysis, we listened to the audio samples after they were converted and observed the presence of noise in the converted samples. Despite experimenting with different models and features, we consistently encountered issues with noise in the converted samples. Additionally, upon examining the spectrograms of the samples, we noticed changes in energy and amplitude, further indicating the presence of unwanted artifacts.

Furthermore, when both the source and target speakers were of the same gender (male or female), we encountered minor difficulties in achieving successful voice conversion.

8 Conclusion

A GMM-based speech conversion system was developed in this project with the aim of transforming a source speaker's voice into that of a target speaker. The project comprised three main components: feature extraction, GMM-based mapping of source and target features, and reconstruction. To assess the effectiveness of the system, we conducted a qualitative evaluation.

Our findings indicate that the GMM-based system struggled to generate high-quality converted speech samples that closely resembled the speech of the intended target speaker. Noise and artifacts were prevalent in the converted samples, and we observed changes in energy and amplitude when examining spectrograms. Despite these challenges, we were able to successfully insert the transformed audio between the source and target speech.

While our system showed promise, further research and refinement are necessary to improve the quality of voice conversion and address the identified issues.

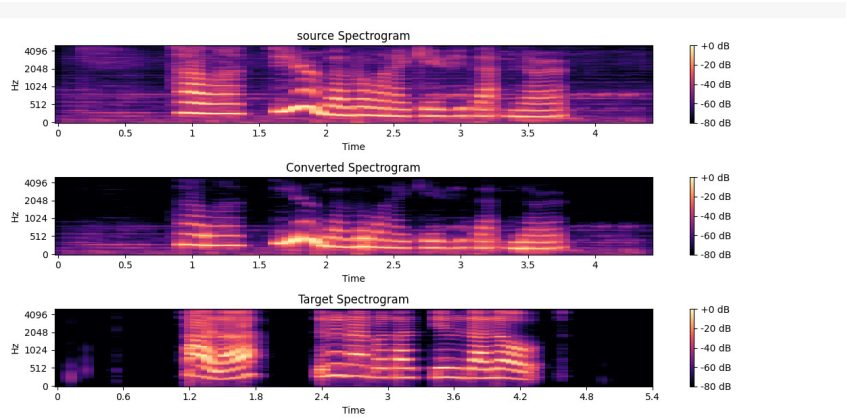

Figure 2: Spectrogram Comparison: Source (left), Converted (middle), Target (right)

This comprehensive methodology represents a blend of sophisticated statistical modeling and a nuanced understanding of speech dynamics. Through this approach, we aim to achieve high-quality voice conversions that are both natural and convincing.

9 References

1. A. W. Black and P. Taylor, "The Festival Speech Synthesis System," 2009. [Online]. Available: `https://www.cs.cmu.edu/~awb/papers/icassp2009/2008_28.pdf`
2. T. Toda, A. W. Black, and K. Tokuda, "Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory," IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 8, pp. 2222-2235, 2007. [Online]. Available: `https://ieeexplore.ieee.org/document/4317579`
3. "Design principles," nnmnkwii documentation, [Online]. Available: `https://r9y9.github.io/nnmnkwii/v0.0.1/design.html`