Indo Asian Journal of Management and Entrepreneurship (IAJOME)
‖ ISSN: 2319-2992, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 17 | Issue: 07 | July 2026 | www.iajome.com ‖

A Review On Voice Conversion Using Artificial Intelligence And Deep Learning

Dr.A.Somasundaram[1],Harshath.M[2]
[1]Assistant Professor, Department of Computer Applications, Sri Krishna Arts and Science College, Coimbatore

[2,3]Students III BCA-B, Department of Computer Applications, Sri Krishna Arts and Science College, Coimbatore

Abstract - *Voice Conversion (VC) is an important field of speech synthesis that focuses on changing a speaker's voice to sound like another person's voice while preserving the original speech content and meaning. This review summarizes fifteen key research studies that highlight the evolution of voice conversion technology, from traditional statistical approaches such as Gaussian Mixture Models (GMMs) and Conditional Restricted Boltzmann Machines (CRBMs) to advanced deep learning techniques including Sequence-to-Sequence (Seq2Seq) models, Generative Adversarial Networks (GANs), and fully convolutional networks. It also discusses major advancements such as one-shot voice conversion through representation disentanglement, the OpenVoice framework for instant voice style cloning, and raw audio generation methods that remove the need for conventional vocoders. In addition, the review explores the integration of Speech-to-Text (STT) systems and Natural Language Processing (NLP) for emotion analysis in applications such as mental health support and market research. Finally, it examines the performance of voice conversion systems through the Voice Conversion Challenges (VCC) and highlights the role of MOSNet, an automated evaluation model that provides speech quality assessments closely aligned with human judgments.*

I. INTRODUCTION

The Voice conversion (VC) is a speech processing technology that aims to change the unique characteristics of a person's voice so that it sounds like another speaker, while keeping the original words, pronunciation, and meaning unchanged. Unlike speech synthesis, which generates speech from text, voice conversion works by transforming an existing speech signal into the voice of a target speaker. To achieve this, the system modifies speaker-specific features such as the spectral envelope (formants), fundamental frequency (F0), speaking rate, duration, and voice quality, while preserving the linguistic information contained in the speech. A speaker's voice is generally described by three main components: linguistic factors, which represent the words and language being spoken; supra-segmental factors, which include prosody, intonation, rhythm, stress, and pitch; and segmental factors, which refer to the acoustic characteristics of individual speech sounds, such as formants and spectral features. A successful voice conversion system must accurately transform both the supra-segmental and segmental characteristics of the source speech to closely match the target speaker while maintaining speech clarity, intelligibility, and naturalness. Recent advances in Artificial Intelligence (AI) and Machine Learning (ML), particularly deep learning techniques, have significantly improved the performance of voice conversion systems by enabling more realistic and expressive speech generation with minimal training data. At the same time, the rapid development of Speech-to-Text (STT) technology has expanded the applications of voice processing beyond speaker identity conversion. Modern STT systems can accurately convert spoken language into text, which can then be analyzed using Natural Language Processing (NLP) and emotion recognition techniques to identify the speaker's emotional state based on both speech patterns and textual content. This combination of voice conversion, STT, and AI-powered emotion analysis has opened new opportunities in areas such as virtual assistants, customer service, healthcare, mental health monitoring, education, accessibility, and market research.

Fig 1. An Assistant for the convertion of voice using Ai

Indo Asian Journal of Management and Entrepreneurship (IAJOME)
‖ ISSN: 2319-2992, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 17 | Issue: 07 | July 2026 | www.iajome.com ‖

II. Fundamentals of Voice Conversion Frameworks

The The standard architecture of a voice conversion system is commonly based on an **analysis–mapping–reconstruction** pipeline, where each stage plays a vital role in transforming the source speaker's voice into the target speaker's voice while preserving the original speech content. In the **analysis** stage, the input speech is first processed to extract important acoustic features such as the spectral envelope, fundamental frequency (F0), energy, duration, and other speech characteristics that describe both the linguistic content and the speaker's unique vocal identity. These extracted features serve as the foundation for the conversion process. In the **mapping** stage, advanced algorithms or machine learning models learn the relationship between the source speaker's features and the target speaker's features. Traditional voice conversion systems used statistical models such as Gaussian Mixture Models (GMMs), while modern approaches rely on deep learning techniques, including neural networks, Generative Adversarial Networks (GANs), and Sequence-to-Sequence (Seq2Seq) models, to perform more accurate and natural feature transformation. The objective of this stage is to modify speaker-specific characteristics while preserving the original pronunciation, linguistic information, and speaking style

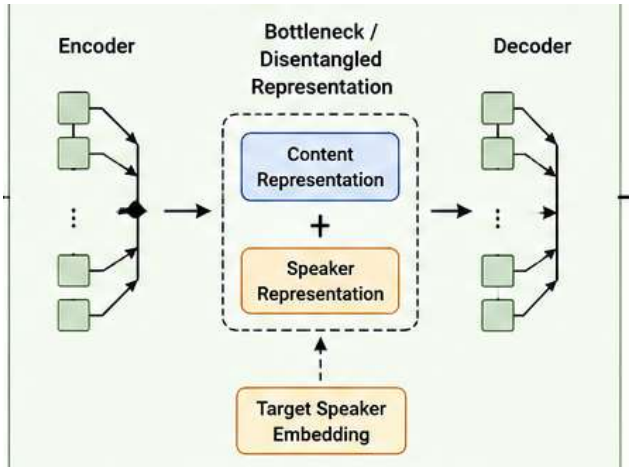

Fig 2. AI-Model of the voice convertion working

2.1 Pipeline Components

Speech analysis is the first stage of the voice conversion process, where the input speech signal is analyzed to extract important acoustic information such as pitch, energy, and spectral features. The speech is converted into intermediate representations using techniques like PSOLA, Harmonic plus Noise Models (HNM), STRAIGHT, or WORLD. This step helps separate speaker-specific characteristics from the actual speech content, making the conversion process more accurate.

FeatureExtraction:
In this stage, the analyzed speech is converted into compact acoustic features that represent the unique characteristics of the speaker. Common features include **Mel-Cepstral Coefficients (MCEP)**, **Line Spectral Frequencies (LSF)**, **Mel-Frequency Cepstral Coefficients (MFCC)**, and **fundamental frequency (F0)**. These low-dimensional features reduce computational complexity while preserving important voice information.

FeatureMapping:
Feature mapping is the core stage of the voice conversion system. Here, the extracted features of the source speaker are transformed to match those of the target speaker using statistical or deep learning models. The goal is to modify the speaker's voice characteristics while preserving the original words, pronunciation, and meaning of the speech.

Reconstruction:
The final stage is reconstruction, where the converted features are transformed back into a natural speech signal using a **vocoder**. Traditional vocoders such as **WORLD** and **STRAIGHT**, along with modern neural vocoders like **WaveNet** and **HiFi-GAN**, generate clear, natural, and high-quality speech that closely resembles the target speaker's voice.

2.2 Statistical Baseline: The GMM and sprocket

In the early development of voice conversion technology, **Gaussian Mixture Models (GMMs)** were considered one of the most effective statistical approaches for converting a source speaker's voice into a target speaker's voice. GMMs learn the relationship between the acoustic features of both speakers by using **parallel speech data**, where the source and target speakers record the same set of sentences. Once this relationship is learned, the model transforms the source speaker's spectral features to match those of the target speaker while preserving the original speech content. Although GMM-based voice conversion achieved reasonable speaker similarity, it often produced **over-smoothed** speech, causing the converted voice to sound

Indo Asian Journal of Management and Entrepreneurship (IAJOME)
‖ ISSN: 2319-2992, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 17 | Issue: 07 | July 2026 | www.iajome.com ‖

unnatural, muffled, or robotic due to the loss of fine acoustic details. To support research and provide a common implementation of these methods, the **Sprocket** software toolkit was introduced as an open-source voice conversion framework. It incorporates **Dynamic Time Warping (DTW)** to accurately align speech frames between the source and target speakers and **Maximum Likelihood Parameter Generation (MLPG)** to generate smoother and more stable acoustic parameters during speech synthesis. Owing to its simplicity, reliability, and ease of use, Sprocket became a popular baseline system and was widely adopted in the **Voice Conversion Challenge (VCC) 2018** for evaluating and comparing the performance of newly developed voice conversion techniques.

III. Advanced Deep Learning and Sequence Modeling

Artificial The introduction of deep learning brought a significant advancement in voice conversion technology, changing the way speech transformation is performed. Earlier voice conversion methods mainly relied on statistical models that processed speech one frame at a time, known as frame-level mapping. In this approach, each small segment of the speech signal was converted independently, without considering the relationship between neighboring speech frames. As a result, these methods often produced speech that sounded over-smoothed, robotic, and lacked natural speech flow. They also struggled to capture important speech characteristics such as rhythm, intonation, speaking style, and long-term dependencies, which are essential for generating realistic speech.

Deep learning transformed voice conversion by treating it as a **sequence-level generative task**, where the entire speech sequence is processed as a whole rather than converting individual frames separately. This allows the model to learn both short-term and long-term relationships between speech features, enabling it to preserve the original linguistic content while accurately transforming the speaker's identity. By analyzing complete speech sequences, deep learning models can better capture pronunciation patterns, pitch variations, speaking rate, emotional expression, and voice quality, resulting in speech that sounds much more natural and human-like.

Several advanced deep learning architectures have played an important role in this transformation. **Sequence-to-Sequence (Seq2Seq)** models learn the mapping between source and target speech sequences, making voice conversion more flexible. **Variational Autoencoders (VAEs)** can learn speaker-independent representations, allowing voice conversion even when parallel training data is unavailable. **Generative Adversarial Networks (GANs)** improve the realism of converted speech by using two competing neural networks to generate high-quality voice samples. More recently, **Transformer-based models** have further enhanced voice conversion by capturing long-range dependencies within speech and producing highly accurate and expressive voice transformations.

Compared with traditional statistical approaches, deep learning-based voice conversion systems achieve higher speaker similarity, better speech quality, and greater naturalness. They require less manual feature engineering, can work with limited or non-parallel training data, and are capable of generating expressive speech that closely matches the target speaker. These improvements have made deep learning the foundation of modern voice conversion systems, enabling practical applications in virtual assistants, speech synthesis, voice cloning, accessibility tools, entertainment, multilingual communication, and personalized AI-based voice services.

3.1 High-Order Feature Discovery with CRBMs

To better capture the complex and non-linear characteristics of human speech, researchers introduced **Conditional Restricted Boltzmann Machines (CRBMs)** as an advanced approach for voice conversion. Unlike traditional **Gaussian Mixture Models (GMMs)**, CRBMs can learn more detailed relationships between the source and target speakers by modeling speech as time-series data. During training, speaker-dependent CRBMs automatically identify hidden patterns and important speech features that represent the unique characteristics of a speaker's voice while reducing the influence of phoneme-related variations. This enables the model to preserve the linguistic content of the speech while producing a voice that more closely resembles the target speaker.

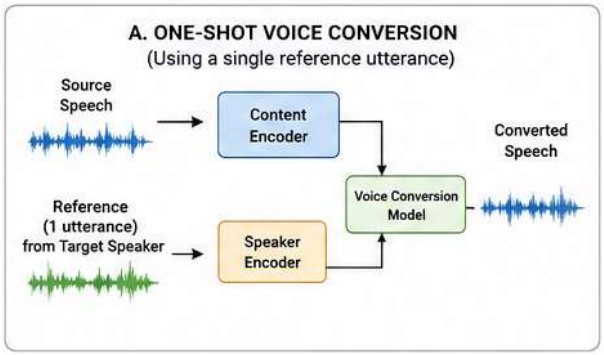

Fig 3. Reference of One-Shot voice conversion

Indo Asian Journal of Management and Entrepreneurship (IAJOME)
‖ ISSN: 2319-2992, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 17 | Issue: 07 | July 2026 | www.iajome.com ‖

3.2 Sequence-to-Sequence (Seq2seq) and Attention

Knitted One of the major limitations of traditional voice conversion methods is their inability to accurately convert the **duration** and **rhythm** of speech. Most conventional systems preserve the timing and speaking pace of the source speaker, which makes the converted speech sound less natural and reduces its similarity to the target speaker. To overcome this limitation, advanced sequence-to-sequence voice conversion models such as **Sequence-to-sequence ConvErsion NeTwork (SCENT)** and **ConvS2S-VC** were developed. These models use **attention mechanisms** to automatically align the source and target speech sequences, allowing them to learn differences in speech duration, rhythm, and pronunciation without requiring manual alignment. **SCENT** employs a pyramid Bidirectional Long Short-Term Memory (BLSTM) encoder and an auto-regressive decoder to simultaneously predict **Mel-spectrograms** and speech durations. It also incorporates bottleneck features extracted from an Automatic Speech Recognition (ASR) model, which helps improve pronunciation accuracy and speech intelligibility. In contrast, **ConvS2S-VC** uses a fully convolutional architecture that enables faster and more efficient processing on GPUs.

IV. One-Shot, Cross-Lingual, and Style-Cloning Innovations

Modern Voice Conversion (VC) research focuses on improving the flexibility and practicality of voice conversion systems by reducing the amount of data required and enabling conversion across different languages. Earlier VC methods depended on large speech datasets from both the source and target speakers, making them difficult and expensive to use. To overcome this limitation, researchers developed **one-shot voice conversion**, where the system can learn the vocal characteristics of a target speaker from just a single speech sample. This allows the model to capture features such as pitch, tone, speaking style, and speaker identity while preserving the linguistic content of the source speech. An even more advanced approach, known as **zero-shot voice conversion**, enables the system to convert speech into the voice of a speaker it has never seen during training by using speaker embeddings extracted from a short reference recording These systems separate speaker identity from language-dependent speech content, making it possible to preserve the original message while generating speech that sounds like the target speaker, regardless of the language being spoken. Recent advances in self-supervised learning models, such as HuBERT and Wav2Vec 2.0, have further improved voice conversion by learning powerful speech representations from large amounts of unlabeled audio, leading to better naturalness, pronunciation, and speaker similarity even with limited training data.

4.1 Representation Disentanglement

The One-shot voice conversion is an advanced technique that enables a voice conversion system to replicate the voice of a target speaker using only a single speech sample, even if that speaker was never included during the model's training. Unlike traditional methods that require large amounts of target speech data, one-shot VC learns the unique vocal characteristics, such as pitch, tone, speaking style, and voice quality, from just one utterance. The main objective is to preserve the linguistic content of the source speech while generating output that closely matches the target speaker's voice. To achieve this, researchers have proposed several effective techniques. **Chou et al.** introduced **Instance Normalization (IN)** in the content encoder to remove speaker-specific information from the speech features. By eliminating global voice characteristics, the encoder is forced to learn only the linguistic content, making it easier to combine this content with the target speaker's voice. Another important contribution was made by **Lu et al.**, who proposed the use of **Global Speaker Embeddings (GSEs)** together with a conditional **WaveNet** speech synthesizer. The extracted speaker embedding is then combined with the content features to generate natural-sounding speech in the target speaker's voice. Since the model can obtain speaker information directly from a single speech sample, it supports **any-to-any voice conversion** without requiring additional adaptation or fine-tuning for new speakers

4.2 The OpenVoice Framework

The The **OpenVoice** framework is a modern voice cloning system designed to achieve instant and high-quality voice cloning with only a short reference speech sample from the target speaker. Unlike traditional voice cloning methods, which often require large amounts of training data and lengthy fine-tuning, OpenVoice adopts a **decoupled architecture** that separates speaker identity from speech style. This design enables the model to generate natural and personalized speech more efficiently. The framework consists of two main components: a **Base Speaker Text-to-Speech (TTS) model** and a **Tone Color Converter**. The Base Speaker TTS model is responsible for generating speech from text while controlling style-related characteristics such as **emotion, accent, speaking rate, rhythm, and intonation**. These features determine how the speech is expressed but do not define the speaker's unique voice. The second component, the Tone Color Converter, transfers the **voice identity** of the target speaker by extracting vocal characteristics such as

Indo Asian Journal of Management and Entrepreneurship (IAJOME)
‖ ISSN: 2319-2992, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 17 | Issue: 07 | July 2026 | www.iajome.com ‖

timbre, pitch, and tone color from a short reference recording. It then applies these characteristics to the speech generated by the Base Speaker TTS model, producing output that sounds like the target speaker while preserving the desired speaking style. Because speaker identity and speech style are learned separately, users can independently modify or combine different voice attributes without retraining the model..

4.3 Raw Audio Conversion: WaveVC

Traditional voice conversion (VC) systems usually generate an intermediate **Mel-spectrogram** from the input speech and then rely on a separate neural **vocoder**, such as **HiFi-GAN**, to convert the Mel-spectrogram into the final speech waveform. Although this two-stage approach has achieved good speech quality, it often suffers from **train-test mismatch**, where the Mel-spectrograms generated during voice conversion differ from the natural spectrograms used to train the vocoder. As a result, the vocoder may introduce unwanted noise, distortions, or unnatural artifacts into the final speech, reducing overall audio quality and speaker similarity. To address these limitations, researchers proposed **WaveVC**, an end-to-end voice conversion framework that performs voice conversion directly on **raw audio waveforms** instead of first generating Mel-spectrograms. By eliminating the separate vocoder stage, WaveVC simplifies the conversion process and reduces the risk of quality degradation caused by mismatched intermediate representations. The model is built using **1D convolutional neural networks**, which process the raw speech waveform directly and learn both speech representation and waveform generation within a single framework. To ensure that the converted speech remains accurate and natural, WaveVC introduces two important training objectives: **Speech Consistency Loss** and **F0 (fundamental frequency) Consistency Loss**

V. HUMAN EMOTIONAL ANALYSIS VIA SPEECH-TO-TEXT

The Artificial Intelligence (AI) and Machine Learning (ML) are increasingly being used to understand and analyze the emotions expressed in human speech. Instead of focusing only on the spoken words, modern AI systems examine both the linguistic content and the way the words are spoken to identify emotions such as happiness, sadness, anger, fear, excitement, or frustration. The process typically begins with **Speech-to-Text (STT)** technology, where speech is converted into written text using advanced STT engines such as **Google Cloud Speech-to-Text** or **IBM Watson Speech to Text**. Once the speech has been transcribed, **Natural Language Processing (NLP)** techniques are applied to analyze the text and identify emotional cues based on word choice, sentence structure, context, and speech patterns. In addition to textual information, some systems also consider vocal characteristics such as pitch, tone, speaking speed, pauses, and intonation to improve emotion recognition accuracy. One of the most widely used techniques in this area is **Sentiment Analysis**, where deep learning models, particularly **Long Short-Term Memory (LSTM)** networks, analyze the sequence of words in a sentence to classify the speaker's emotional state as **positive**, **negative**, or **neutral**. Since LSTM models can learn long-term relationships between words, they are highly effective at understanding the context of conversations and detecting emotions even in complex sentences. Beyond analyzing individual sentences, modern AI systems also perform **Contextual Analysis**, which monitors how emotions change throughout an entire conversation rather than evaluating each sentence independently. By tracking the emotional progression over time, these systems can detect changes in mood, identify signs of stress or anxiety, and provide a more accurate understanding of the speaker's emotional state. This capability is particularly valuable in **mental health monitoring**, where continuous emotion tracking can help identify early symptoms of depression, anxiety, or emotional distress, allowing healthcare professionals to provide timely support.

VI. BENCHMARKING AND OBJECTIVE ASSESSMENT

The **Voice Conversion Challenge (VCC)** is a globally recognized benchmark event that has played a significant role in advancing voice conversion research by providing a common platform for evaluating and comparing different voice conversion technologies. Organized every two years, the challenge brings together researchers, academic institutions, and industry experts from around the world to develop and test their latest voice conversion systems using the same datasets, evaluation methods, and performance metrics. Before the introduction of the VCC, it was difficult to fairly compare different voice conversion approaches because researchers used different speech datasets, experimental settings, and evaluation procedures. The VCC solved this problem by offering a standardized evaluation framework, allowing all participating systems to be tested under identical conditions. Participants are provided with carefully prepared speech datasets containing recordings from multiple speakers and are asked to convert speech from one speaker to another while preserving the original linguistic content and producing natural, high-quality speech that closely matches the target speaker's voice. The converted speech is then evaluated using both **objective measures**, such

Indo Asian Journal of Management and Entrepreneurship (IAJOME)
‖ ISSN: 2319-2992, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 17 | Issue: 07 | July 2026 | www.iajome.com ‖

as speaker similarity and speech quality metrics, and **subjective listening tests**, where human listeners rate the naturalness, intelligibility, and similarity of the converted speech to the target speaker

6.1 Lessons from VCC 2018 and 2020

The **Voice Conversion Challenge (VCC) 2018** represented a significant milestone in the advancement of voice conversion technology by encouraging researchers to develop **non-parallel voice conversion** systems through its **"Spoke" task**. Unlike traditional voice conversion methods that required parallel speech datasets containing the same sentences spoken by both the source and target speakers, the Spoke task challenged participants to build models that could perform voice conversion without relying on such paired data. This made the technology much more practical because collecting parallel recordings is costly, time-consuming, and often impossible in real-world scenarios. The challenge accelerated the adoption of advanced deep learning techniques, particularly **WaveNet-based neural vocoders**, which were capable of generating highly realistic speech waveforms. Among the participating systems, **Team N10** achieved some of the best results by producing speech with improved naturalness, speaker similarity, and intelligibility. Nevertheless, the evaluation results also revealed that even the best-performing systems had not yet reached the quality of natural human speech. Human listeners could still identify differences in speech smoothness, pronunciation, and overall audio quality, indicating that further improvements were needed in speech modeling and waveform generation. Building on the success of VCC 2018, the **Voice Conversion Challenge (VCC) 2020** focused on the more challenging problem of **cross-lingual voice conversion**, where the source and target speakers communicate in different languages. This task required voice conversion systems not only to preserve the speaker's unique vocal identity but also to accurately convert speech across languages with different phonetic structures, pronunciation rules, and speaking styles. The results showed that modern deep learning models had made remarkable progress in **intra-lingual voice conversion**, with several systems achieving speaker similarity that was close to human performance when both speakers used the same language. However, the challenge also exposed the limitations of existing methods in cross-lingual settings. Researchers identified a major obstacle known as the **"language barrier,"** where language-specific characteristics present in the target speaker's reference speech influenced the quality of the converted output. As a result, the converted speech often contained unnatural accents, pronunciation inconsistencies, reduced fluency, and lower speaker similarity when the source and target speakers belonged to different languages

6.2 Automated Evaluation: MOSNet

Evaluating the quality of voice conversion (VC) systems is an essential part of voice conversion research, as it helps determine how natural and realistic the converted speech sounds. Traditionally, the quality of converted speech is measured using **Mean Opinion Score (MOS)**, where a group of human listeners rates the naturalness of speech on a scale, typically from **1 (poor quality) to 5 (excellent quality)**. Although MOS is considered the most reliable method because it reflects human perception, conducting these subjective listening tests is expensive, time-consuming, and labor-intensive. Researchers must recruit a large number of listeners, prepare listening experiments, collect ratings, and analyze the results. Furthermore, the ratings may vary depending on the listeners' backgrounds, hearing abilities, and personal preferences, making the evaluation process difficult to repeat consistently. To overcome these challenges, researchers developed **MOSNet**, a deep learning-based objective speech quality assessment model that automatically predicts human MOS ratings without requiring listening tests. MOSNet is trained using large datasets containing speech samples along with their corresponding human-assigned MOS scores, enabling the model to learn the relationship between speech characteristics and human perception of quality. The architecture of MOSNet combines **Convolutional Neural Networks (CNNs)** and **Bidirectional Long Short-Term Memory (BLSTM)** networks to analyze speech effectively. The CNN layers extract important local acoustic features, such as spectral patterns and speech characteristics, while the BLSTM layers capture long-term temporal relationships and contextual information across the entire speech sequence. By combining these two deep learning techniques, MOSNet can accurately estimate how natural and high-quality a speech sample would sound to human listeners. Experimental results have shown that MOSNet achieves a **high correlation with human Mean Opinion Scores**, particularly at the **system level**, meaning that it can reliably rank and compare different voice conversion systems in a way that closely matches human evaluations.

VII . Conclusion

Several The field of voice conversion and speech-based emotional analysis has experienced remarkable growth over the past decade, evolving from traditional signal processing techniques to advanced deep learning models capable of producing highly natural and expressive speech. Early voice conversion systems mainly relied on simple spectral mapping methods, which focused

Indo Asian Journal of Management and Entrepreneurship (IAJOME)
‖ ISSN: 2319-2992, Impact Factor 5.007 ‖
‖ Peer-Reviewed | Open Access | International Journal ‖
‖ Volume: 17 | Issue: 07 | July 2026 | www.iajome.com ‖

on modifying the acoustic characteristics of speech to match a target speaker. Although these approaches could change the speaker's voice, they often produced speech that sounded robotic, lacked naturalness, and required large amounts of parallel training data. With the rapid advancement of Artificial Intelligence (AI) and deep learning, researchers introduced more powerful techniques such as sequence-to-sequence attention models, which can automatically learn the relationship between source and target speech while simultaneously modeling speech duration, pronunciation, and prosody. Another major breakthrough has been representation disentanglement, where speaker identity, linguistic content, and speech style are learned as separate components. This enables voice conversion systems to preserve the spoken message while independently modifying the speaker's voice, emotional expression, or speaking style. Furthermore, decoupled voice cloning frameworks, such as OpenVoice, separate speaker identity from speech style, making it possible to perform high-quality one-shot voice cloning using only a single reference speech sample. These innovations have significantly reduced the amount of training data required while improving the naturalness, flexibility, and scalability of modern voice conversion systems.

At the same time, significant progress has been made in **speech-based emotional analysis**, where AI systems combine **Speech-to-Text (STT)** technology with **Natural Language Processing (NLP)** and deep learning models to recognize and interpret human emotions from spoken conversations. Modern systems not only analyze the words spoken but also examine vocal features such as pitch, tone, speaking rate, rhythm, pauses, and intonation to detect emotions such as happiness, sadness, anger, fear, or excitement. Advances in **raw audio generation** have further enhanced speech synthesis by enabling models to generate speech directly from audio waveforms instead of relying on intermediate representations like Mel-spectrograms. This approach reduces speech distortion, improves voice quality, and creates more realistic and expressive synthetic speech. Together, these developments are moving AI closer to replicating the complexity of human communication by producing speech that not only sounds like a specific person but also conveys appropriate emotions, speaking styles, and expressive characteristics.

REFERENCES:

1. **Sisman, B., et al.** "An Overview of Voice Conversion and Its Challenges: From Statistical Modeling to Deep Learning."
2. **Nakashika, T., et al.** "Voice conversion using speaker-dependent conditional restricted Boltzmann machine."
3. **MyShell AI.** "OpenVoice: Instant Voice Cloning."
4. **Shukla, A.** "Utilizing AI and Machine Learning for Human Emotional Analysis through Speech-to-Text Engine Data Conversion."
5. **Lu, H., et al.** "One-shot Voice Conversion with Global Speaker Embeddings."
6. **Chou, J., et al.** "One-shot Voice Conversion by Separating Speaker and Content Representations with Instance Normalization."
7. **Kobayashi, K. & Toda, T.** "sprocket: Open-Source Voice Conversion Software."
8. **Lorenzo-Trueba, J., et al.** "The Voice Conversion Challenge 2018: Promoting Development of Parallel and Nonparallel Methods."
9. **Kameoka, H., et al.** "ConvS2S-VC: Fully Convolutional Sequence-to-Sequence Voice Conversion."
10. **Zhang, J., et al.** "Sequence-to-Sequence Acoustic Modeling for Voice Conversion."
11. **Yi, Z., et al.** "Voice Conversion Challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion."
12. **Mohammadi, S.H. & Kain, A.** "An overview of voice conversion systems."

13. **Lo, C., et al.** "MOSNet: Deep Learning-based Objective Assessment for Voice Conversion."

14. **Ko, K., et al.** "WaveVC: Speech and Fundamental Frequency Consistent Raw Audio Voice Conversion."