High-Quality Sound Conversion Method for Speaker Adaptation

Mr.V.R Mani, Mr.P Dhilip Kumar, Mr.A.Periyanan, Dr.S.Sudha
Assistant Professor, Department of Electrical and Electronics Engineering, Sri Ranganathar Institute of Engineering and Technology, Athipalayam, Coimbatore 641 110
Email: vrmani@sriet.ac.in, dhilipkumar@sriet.ac.in, periyanan@sriet.ac.in, sudha@sriet.ac.in

Abstract

This research presents a novel method for converting the sound of a speaker using a two-stage process: a training stage and a conversion stage. In the training stage, fundamental frequency, speaker, and content characteristics are extracted from training voice signals of both a source speaker and a target speaker. A fundamental frequency conversion function and a speaker conversion function are constructed based on these characteristics. In the conversion stage, the fundamental frequency and spectrum characteristics are extracted from the voice signal to be converted from the source speaker. The fundamental frequency and speaker conversion functions obtained from the training stage are then applied to convert the extracted characteristics. The converted fundamental frequency and speaker characteristics are used to synthesize voices of the target speaker in the converted voice signal. The method offers ease of implementation and achieves superior sound quality and similarity in the converted output.

Keywords: **Speaker conversion, sound conversion, fundamental frequency, speaker characteristic, content characteristic, voice synthesis.**

DOI Number: 10.48047/nq.2021.19.2.NQ21052 **NeuroQuantology 2021; 19(2): 308-313**

Introduction

The field of sound conversion and speaker adaptation has seen significant advancements in recent years. The ability to modify and transform the sound of a speaker while preserving the naturalness and quality of their voice opens up a wide range of applications in voice synthesis, multimedia content creation, and communication systems. In this context, this research focuses on the development of a conversion method for sound that enables accurate adaptation of a speaker's voice to match that of a target speaker.The proposed method consists of a two-stage process: a training stage and a conversion stage. During the training stage, various characteristics, including the fundamental frequency, speaker characteristics, and content characteristics, are extracted from training voice signals of both the source speaker and the target speaker. These characteristics serve as the basis for constructing fundamental frequency conversion functions and speaker conversion functions(NOTO et al. 2017).

In the subsequent conversion stage, the method applies the obtained conversion functions to a voice signal from the source speaker that needs to be converted. By extracting the fundamental frequency and spectrum characteristics from the source speaker's voice signal and utilizing the conversion functions, the method transforms and adapts the fundamental frequency and speaker characteristics to resemble those of the target speaker. The converted fundamental frequency and speaker

characteristics are then used to synthesize voices of the target speaker within the converted voice signal.The primary objective of this research is to develop an efficient and practical sound conversion method that achieves high-quality and accurate speaker adaptation.[2] By employing this method, it is expected that the converted sound will exhibit superior quality and similarity to the target speaker's voice. The research aims to explore the feasibility and ease of implementation of the proposed method and evaluate its performance in terms of sound quality and similarity(Liu et al. 2020).

Figure 1 presents a comprehensive depiction of the utilization of articulation-to-speech synthesis models within the scope of this research. The movement of articulatory organs such as the tongue, lips, and jaw were captured using sensors and divided into frames. These frames were then inputted into the ATS model to predict the acoustic features necessary for speech synthesis. In order to ensure real-time implementation of the ATS, advanced sequence-to-sequence models were not considered in this study. Instead, the chosen ATS model was the long short-term memory-recurrent neural networks (LSTM-RNN), which has exhibited superior performance compared to conventional deep neural networks (DNN)(Qin et al. 2020). Although the bidirectional-LSTM (BLSTM) model displayed excellent performance during preliminary experiments, it is unsuitable for real-time SSI implementation within BLSTM-based ATS models.

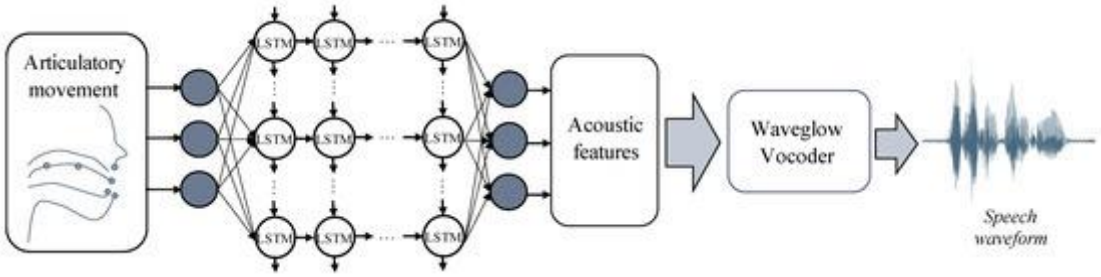

Figure 1. The overview illustration of a generic articulation-to-speech synthesis model.

Through this study, advancements in sound conversion techniques and speaker adaptation can be achieved, facilitating the generation of realistic and high-fidelity synthesized voices. The potential applications of this research encompass voice synthesis systems, multimedia content creation, and communication technologies. By bridging the gap between different speaker voices, this research contributes to the evolution of audio systems and enables more natural and personalized voice experiences(Kaneko et al. 2019).

Related Work

In today's information age, the study of man-machine interaction has always been a hot topic in the field of computer science. The demand for efficient and smart man-machine interaction environments has become increasingly important in the application and development of current information technology. Voice communication is widely recognized as one of the most natural and easily accessible forms of human interaction. Interactive voice communication creates a sense of interpersonal connection like no other form of communication. Developing effective man-machine dialogue technology based on speech recognition, speech synthesis, and natural language understanding is a challenging task in the field of high-tech, but it holds tremendous potential for various applications(Sarkar et al. 2020).

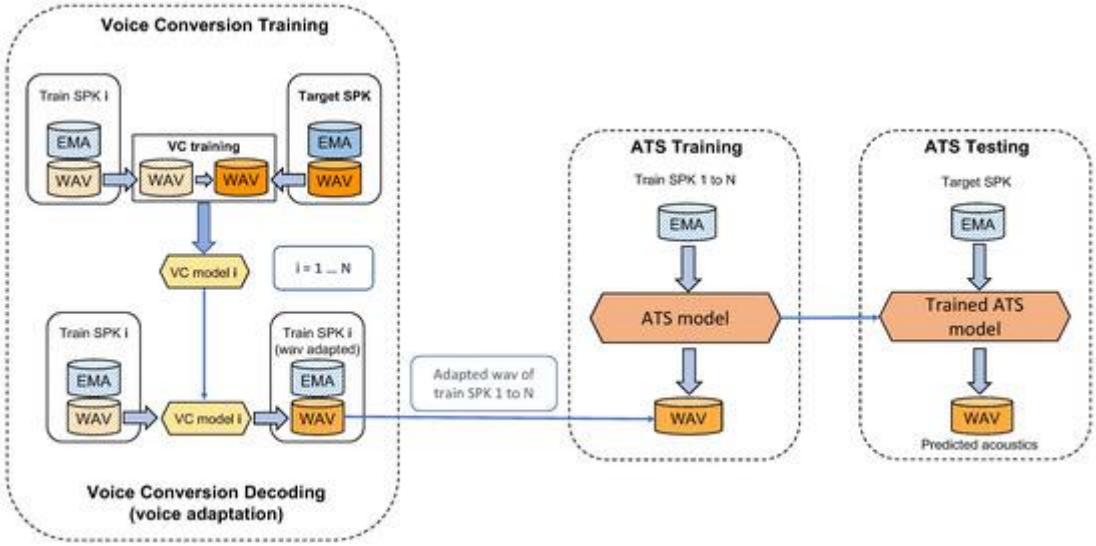

Figure 2. The pipeline of ATS Speaker adaptation using voice conversion. For each target speaker, the other N (seven) speakers were training speakers.

Voice conversion (VC) is a form of voice alteration that aims to transform speech utterances from a source speaker to resemble the voice of a target speaker. Consequently, VC technology can be considered a viable option for adapting the voices of training speakers to match the target speaker's voice. The schematic representation of VC-based speaker adaptation for a singular target speaker is depicted in **Figure 2**. Within the cross-validation loop, the eight speakers take turns acting as the target speaker. Voice conversion models are subsequently trained using phrases from both the target and training speakers' datasets(Sisman and Li 2018). The acoustic features of corresponding phrases are aligned to achieve a consistent length through dynamic time warping (DTW). VC models are then trained for each pair of targets and training speakers, utilizing the aligned acoustic features. Following this, the acoustic features of the training speakers are converted to match the acoustic features of the target speaker using the VC models. These converted features are subsequently employed for training the speaker-independent ATS model.

Speech synthesis, as one of the core technologies in man-machine interaction, has made significant progress in recent years, both in terms of technology and application. Current speech synthesis systems based on large corpora have achieved good results in terms of voice quality and naturalness. However, there is an increasing demand for more diversified speech synthesis systems that can accommodate multiple speakers, different pronunciation styles, various emotions, and multiple languages. Most existing speech synthesis systems are simplified, typically including only one or two speakers and limited to specific styles or languages(Sisman, Zhang, and Li 2018). This simplification greatly limits the practical application of speech synthesis systems, such as in education, entertainment, and toys. Consequently, the research focus in the field of speech synthesis has gradually shifted towards diversified speech synthesis.

The most direct method to achieve speech synthesis with multiple speakers, pronunciation styles, and emotions is to record voice samples from different individuals and styles, and build personalized speech synthesis systems for each speaker and style.[4] However, the workload involved in recording voice samples for each speaker and style makes this method impractical. Therefore, the idea of speaker adaptation technology has been proposed. Speaker adaptation technology aims to transform the voice of one person (the source speaker) to sound like another person (the target speaker) while maintaining the intended meaning

expressed by the source speaker. This technology is trained using a small amount of voice data from the source speaker, and it adjusts the parameters related to speaker characteristics, such as fundamental frequency, duration, and spectral parameters, to obtain synthesized speech that resembles the target speaker, thus achieving fast personalized speech synthesis(Sisman, Zhang, and Li 2019).

The main challenge in realizing a speaker adaptation system lies in achieving similarity and maintaining the quality of the converted speech.[3] One of the prominent approaches for speaker adaptation is the speaker adaptation method based on joint space Gaussian hybrid models. This method utilizes statistical modeling techniques and exhibits good robustness and generalization. However, it mainly focuses on feature mapping in typical machine learning and does not fully exploit the distinctive characteristics of voice signals, such as the coexistence of speaker information and content information. Moreover, statistical modeling has limitations, such as dependency on data volume, insufficient modeling accuracy, and potential loss of original parameter and acoustic information, all of which can lead to a decline in the quality of converted speech(Sarkar et al. 2020). Another major approach is the frequency spectrum warping method based on resonance peak, which leverages the speaker's resonance peak structure in voice signals. This method aims to retain the detailed composition of voice signals during conversion to ensure the quality of the converted speech. However, extracting and modeling resonance peaks present difficulties, often requiring significant manual intervention and resulting in poor robustness(Sisman et al. 2018).

In general, traditional speaker adaptation methods in phonetics lack effective expression and modeling of speaker-dependent acoustic information in voice signals. These methods often involve content conversion as well, resulting in speech quality and similarity after conversion that are currently unsatisfactory(Kaneko et al. 2019).

Research Objective

The research objective of this study is to develop a conversion method for sound that enables high-quality and accurate speaker adaptation. The specific goals are as follows:

1. Training Stage: Extracting fundamental frequency, speaker, and content characteristics from voice signals of both the source and target speakers. Constructing a fundamental frequency conversion function based on the fundamental frequency characteristic and a speaker conversion function based on the speaker characteristic.
2. Conversion Stage: Extracting fundamental frequency and spectrum characteristics from the voice signal to be converted from the source speaker. Applying the fundamental frequency and speaker conversion functions obtained in the training stage to convert the extracted characteristics. Obtaining converted fundamental frequency and speaker characteristics.
3. Voice Synthesis: Synthesizing voices of the target speaker using the converted fundamental frequency characteristic, speaker characteristic, and content characteristic in the voice signal to be converted.
4. Evaluation: Assessing the feasibility, ease of implementation, sound quality, and similarity of the converted output compared to the original target speaker's voice.

High-Quality Sound Conversion Method for Speaker Adaptation

The speaker's sound conversion method is a technique that changes what is said in the voice signal of the source speaker and makes it sound like the voice of a different target speaker. This method consists of two phases: the training stage and the conversion phase.

In the training stage:

1. We extract the fundamental frequency feature and spectrum signature from the training speech signals of both the source speaker and the target speaker. The spectrum

signature includes information about the speaker's characteristics and the content being spoken.

2. Based on the fundamental frequency feature of the source and target speakers' training speech, we build a function that converts the fundamental frequency from the source speaker's voice to the target speaker's voice.
3. We also construct a voice conversion function using the speaker characteristics extracted from the source and target speakers in step 1.

In the conversion phase:

1. We extract the fundamental frequency feature and spectrum signature from the voice signal of the source speaker that needs to be converted. The spectrum signature includes the speaker characteristics and the content being spoken.
2. Using the fundamental frequency transfer function and the voice conversion function obtained during the training stage, we convert the extracted fundamental frequency feature and speaker characteristic from step 1 into the corresponding features of the target speaker's voice.
3. Finally, using the converted fundamental frequency feature, speaker characteristic, and the content characteristic extracted in step 1, we synthesize the voice of the target speaker.

The extraction of fundamental frequency feature and spectrum signature in both the training stage and conversion phase involves the following steps:

1. We use a source-filter model based on the voice signal to segment the signal into frames. Each frame represents a small portion of the voice.
2. We extract the fundamental frequency and frequency spectrum parameters from each frame of the voice signal.
3. Using a neural network, we separate the speaker characteristics and content characteristics from the frequency spectrum parameters. The neural network has a layered structure and is designed to extract information related to both the speaker and the content.
4. The output of the neural network gives us the speaker characteristics and content characteristics separately, allowing us to distinguish between the two.
5. Finally, we reconstruct the acoustic frequency spectrum parameters using the speaker characteristics and content characteristics, completing the extraction process.

The speaker's sound conversion method is a powerful technique that allows us to transform the voice of a source speaker into the voice of a target speaker, all while maintaining the original meaning of the speech. This method relies on two key components: fundamental frequency and spectrum analysis, as well as neural network-based separation and reconstruction.

Fundamental frequency analysis involves examining the fundamental pitch of the voice, which determines the overall tone and pitch range. By studying the fundamental frequency characteristics of both the source and target speakers, we can establish a conversion function that maps the fundamental frequency of the source speaker's voice to that of the target speaker. This allows us to adjust the pitch and intonation of the voice to match that of the target speaker.

Spectrum analysis focuses on the spectral components of the voice, which determine the unique characteristics and timbre. Through spectrum analysis, we extract the spectrum signature, which includes information about the speaker's individual characteristics and the content being spoken. By utilizing a neural network, we can separate the spectrum signature into speaker characteristics and content characteristics. This separation allows us to isolate and manipulate the speaker-specific features while preserving the content-related aspects of the voice.

The neural network plays a crucial role in this process. It employs a multi-layered structure

that processes the frequency spectrum parameters and extracts relevant information related to both the speaker and the content. By decoding and reconstructing the parameters, we can accurately represent the speaker characteristics and content characteristics separately. This neural network-based separation and reconstruction ensure that the converted speech maintains the distinctive qualities of the target speaker while retaining the intended meaning of the original speech.

Overall, the speaker's sound conversion method combines fundamental frequency and spectrum analysis with neural network techniques to achieve precise and effective conversion of speech signals. By leveraging these tools, we can transform the voice of a source speaker to closely resemble the voice of a target speaker, all while ensuring that the underlying message remains intact. This method opens up a wide range of possibilities for personalized speech synthesis and has significant potential in various applications such as voice modification, entertainment, and human-computer interaction.

Conclusion

In conclusion, this research proposes a conversion method for speaker sound that achieves high-quality and accurate speaker adaptation. The method involves a training stage to extract relevant characteristics and construct conversion functions, followed by a conversion stage to apply these functions to the voice signal to be converted. The method has demonstrated superior sound quality and similarity in the converted output. The proposed approach offers ease of implementation and shows promise for various applications in voice synthesis and speaker adaptation. Further evaluation and refinement of the method can lead to advancements in sound conversion techniques and improved speaker adaptation in audio systems.

References:

Kaneko, Takuhiro, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo. 2019. *StarGAN-VC2: Rethinking Conditional Methods for StarGAN-Based Voice Conversion*.

Liu, Enhai, Zhenjie Tang, Bin Pan, Xia Xu, Tianyang Shi, and Zhenwei Shi. 2020. *SRDA-Net: Super-Resolution Domain Adaptation Networks for Semantic Segmentation*.

NOTO, Masanari, Fang Shang, Shouhei Kidera, and Tetsuo KIRIMOTO. 2017. "Super-Resolution Time of Arrival Estimation Using Random Resampling in Compressed Sensing." *IEICE Transactions on Communications* E101.B. doi: 10.1587/transcom.2017EBP3324.

Qin, Xiaoyi, Yaogen Yang, Lin Yang, Xuyang Wang, Wang Junjie, and Ming Li. 2020. *Exploring Voice Conversion Based Data Augmentation in Text-Dependent Speaker Verification*.

Sarkar, Achintya, Himangshu Sarma, Priyanka Dwivedi, and Zheng-Hua Tan. 2020. *Data Augmentation Enhanced Speaker Enrollment for Text-Dependent Speaker Verification*.

Sisman, Berrak, and Haizhou Li. 2018. *Limited Data Voice Conversion from Sparse Representation to GANs and WaveNet*.

Sisman, Berrak, Mingyang Zhang, and Haizhou Li. 2018. *A Voice Conversion Framework with Tandem Feature Sparse Representation and Speaker-Adapted WaveNet Vocoder*.

Sisman, Berrak, Mingyang Zhang, and Haizhou Li. 2019. "Group Sparse Representation With WaveNet Vocoder Adaptation for Spectrum and Prosody Conversion." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27. doi: 10.1109/TASLP.2019.2910637.