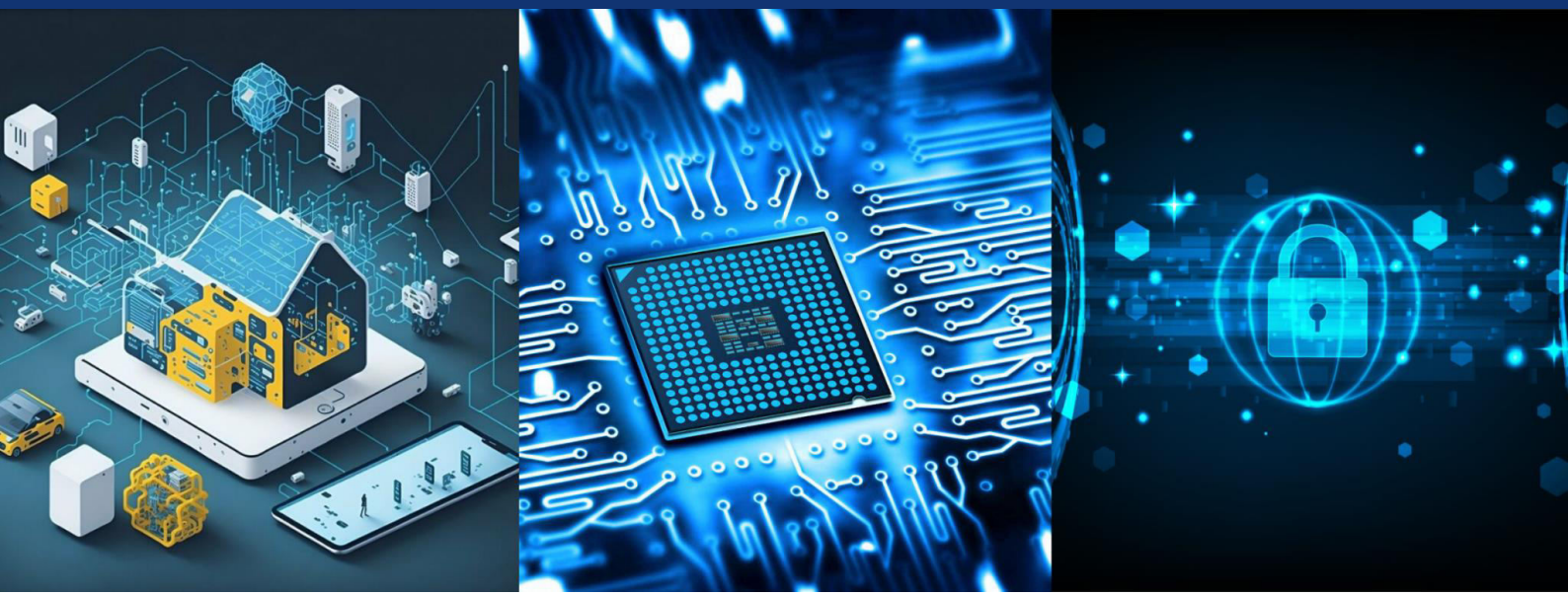
IJIRCCE

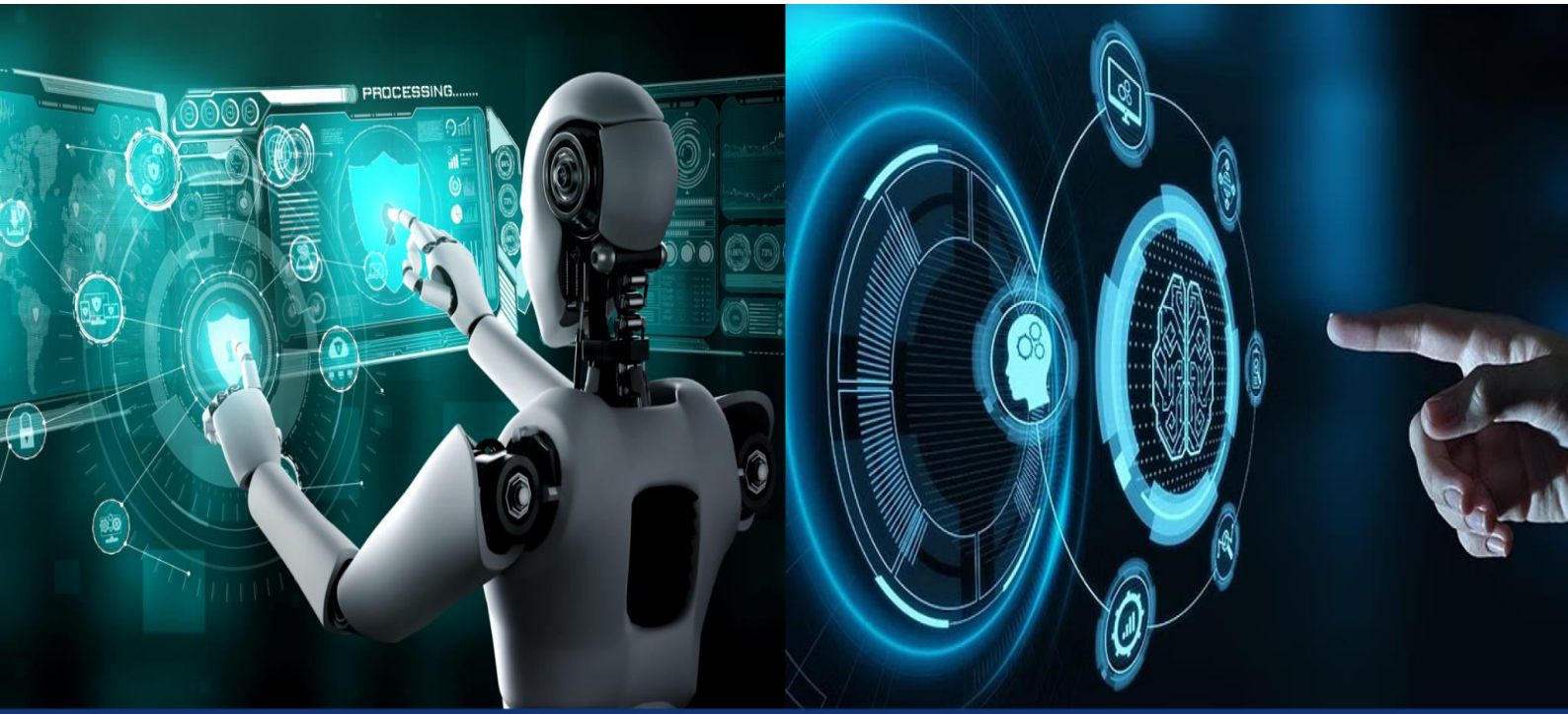
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.771

Volume 13, Issue 4, April 2025

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI:10.15680/IJIRCCE.2025.1304229

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Real-Time Accent Translation

Priyadarshini R, Abhay Bhosle, Ganesh M, Anurup Koli

Assistant professor, Dept. of CSE, Reva University, Bengaluru, Karnataka, India

UG Student, Dept. of CSE, Reva University, Bengaluru, Karnataka, India

UG Student, Dept. of CSE, Reva University, Bengaluru, Karnataka, India

UG Student, Dept. of CSE, Reva University, Bengaluru, Karnataka, India

ABSTRACT: This literature review provides a comprehensive analysis of existing research in the field of accent conversion and its application to real-time speech translation. The focus is on non-autoregressive models, which offer low-latency accent conversion, a critical requirement for real-time applications. The review covers key advancements in phonetic-based accent conversion, zero-shot learning methods, and neural network architectures that enable efficient accent translation. Finally, the review highlights the limitations of traditional methods and emphasizes the role of Nechaev and Kosyakov's non-autoregressive real-time accent conversion model as a foundation for future developments in this area.

KEYWORDS: Accent Conversion, Real-Time Translation, Non-Autoregressive Models, Zero-Shot Learning, Neural Networks

I. INTRODUCTION

Accent translation aims to modify the regional or dialectal characteristics of speech while retaining the speaker's natural voice features. The ability to translate accents in real time is especially valuable in globalized industries such as customer service, education, and international business. Existing methods in accent conversion often struggle with high latency and the requirement for parallel datasets. The development of non-autoregressive models has significantly reduced these limitations, allowing for low- latency, high-quality accent conversion suitable for real-time applications.

The Nechaev and Kosyakov (2024) model forms the basis for this review. Their work introduces a non- autoregressive neural network for real-time accent translation, which processes speech in parallel, thereby addressing the delay issues seen in earlier models. This literature review examines prior research leading to this breakthrough and evaluates current methodologies for accent conversion.

Objectives

The main objectives of this research are:

1. To develop a real-time accent translation system capable of converting live voice between different English accents.
2. To implement a non-autoregressive neural network model that ensures low-latency accent translation.
3. To maintain speaker identity by preserving vocal characteristics such as timbre and pitch during translation.
4. To test the system's accuracy and naturalness in converting between American English and British English accents.

Organization

The rest of the paper is organized as follows:

- Chapter 2 provides a literature review and survey of related work, explaining different traditional approaches used.
- Chapter 3 discusses the system model / methodology, including the feature extraction process, the proposed Neural Network architecture.
- Chapter 4 presents the summery and conclusion of presented research paper.
- Chapter 5 provides references used to create this research paper.

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI:10.15680/IJIRCCE.2025.1304229

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

II. RELATED WORK / LITERATURE SURVEY

2.1 Phonetic-Based Accent Conversion

Traditional methods of accent conversion relied on phonetic frame mapping, where phonetic representations of speech are transformed between accents. Zhao et al. (2019) proposed a phonetic posteriorgram-based frame pairing technique to achieve this conversion. Their model maps phonetic frames between source and target accents, enabling accurate accent shifts while preserving speaker characteristics. However, this approach requires parallel datasets, limiting its scalability to real-world applications.

Aryal and Gutierrez-Osuna (2014) explored the use of voice conversion technologies for accent reduction, improving the intelligibility of non-native speakers. Their approach modifies the acoustic features of speech to make it sound more native-like. Although effective, the reliance on parallel data poses significant challenges for real-time use.

2.2 Zero-Shot Accent Conversion

The limitations of parallel data were addressed through zero-shot learning approaches. Quamer et al. (2022) introduced a zero-shot foreign accent conversion model, enabling accent translation without the need for paired data. This method generalizes well across unseen accents, making it highly adaptable for real-time applications. The model maintains speaker identity while adjusting accent features, offering a flexible solution for environments with diverse accents.

Building on zero-shot learning, Jin et al. (2023) developed a voice-preserving multiple accent conversion model, which further enhances real-time applicability. By utilizing minimal training data, the model achieves low-latency accent conversion while preserving speaker-specific characteristics, addressing the needs of virtual assistants, call centers, and other real-time communication tools.

2.3Neural Network-Based Accent Conversion

Autoregressive vs Non-Autoregressive Models

Autoregressive models like Tacotron 2 predict speech sequentially, making them unsuitable for real-time applications due to their high latency. While Tacotron 2 excels in generating high-quality speech, its sequential nature introduces delays that hinder real-time processing.

Non-autoregressive models, on the other hand, allow for parallel processing of speech frames, significantly reducing processing time. Nechaev and Kosyakov (2024) proposed a non-autoregressive real- time accent conversion model that processes speech frames simultaneously, thus achieving low-latency accent conversion. Their model integrates modules for accent identification, speaker embedding, and gender detection, ensuring that the converted speech retains its naturalness while shifting accents.

Zhouetal. (2023) introduced a **TTS-guided accent conversion model**, which uses a **text-to-speech system** to synthesize speech in a target accent. This approach eliminates the need for parallel data, making it easier to generalize across different accents. The TTS-guided method offers flexibility in real-time applications, enabling efficient and scalable accent translation for global users

III. METHODOLOGY / APPROACH

3.1Overview

The foundation of this project is built upon the non-autoregressive real-time accent conversion model proposed by Nechaev and Kosyakov (2024). The system consists of several key modules:

1. **Speech Recognition (ASR):** Converts input speech into text in real time.
2. **Accent Conversion Module:** Transforms the accent of the recognized text using a non-autoregressive neural network.
3. **Text-to-Speech (TTS) Module**: Converts the accent-modified text back into speech.
4. **Voice Cloning and Speaker Preservation:** Ensures that the speaker's identity and voice characteristics are retained throughout the conversion process.

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI:10.15680/IJIRCCE.2025.1304229

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

3.2 Model Description

The proposed accent conversion method comprises several interconnected modules unified into a single end-to-end architecture. These models are used for accent and gender identification, speaker recognition, speech-to-phonetic token conversion, spectrogram generation, and decoding the resulting spectrogram into an audio signal. Figure 1 shows the overall interaction scheme of these models during the inference stage, which is the generation of the output L1 audio.

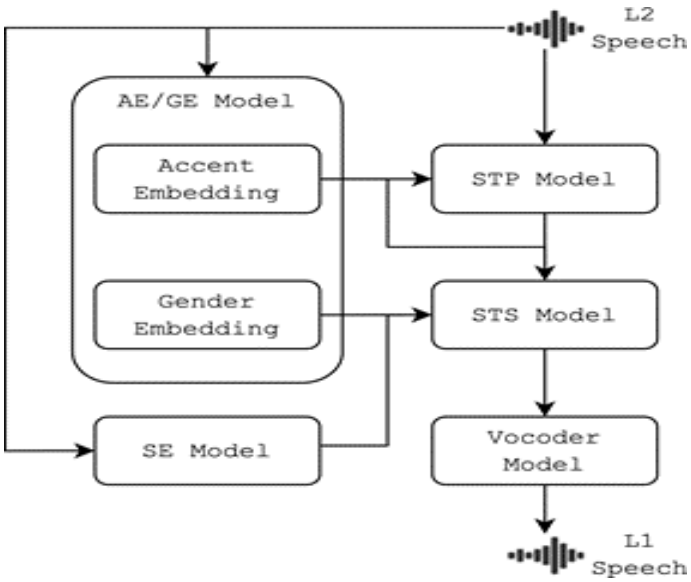


Fig. 1. General scheme of accent conversion model inference with voice cloning.

The incoming audio signal (L2 Speech) is fed into the Speech-to-Phonemes model (STP Model), the Accent Embed-ding and Gender Embedding model (AE/GE Model), and the Speaker Embedding model (SE Model). The accent vector

(Accent Embedding) influences the generation of the phonetic representation, which is then fed into the Speech-to-Speech model (STS Model) to generate the mel- spectrogram. The mel-spectrogram generation is influenced by the same Accent Embedding, the Gender Embedding vector, and the individual vocal characteristics vector output by the SE Model. The generated spectrogram is then converted into the L1 Speech audio signal using the Vocoder Model.
The overall pipeline for generating L1 speech from the original L2 speech can be simplified into the following formula:

$$aL1 = FV\ (FSTS\ (FS(aL2, FAE(aL2)), FAE(aL2), FGE(aL2), FSE(aL2))), (1)$$

where $aL1$ is the generated L1 speech audio signal; $aL2$ is the input L2 speech audio signal; FV – Vocoder Model; $FSTS$ – STS Model; $FSTP$ – STP Model; FAE – AE/GE Model, Accent Embedding; FGE – AE/GE Model, Gender Embedding; FSE – SE Model, Speaker Embedding.
To create a unified end-to-end accent conversion model, each individual model must be trained sequentially. The AE/GE Model and SE Model are independent of other models and can be trained in any order. Training the STP Model re- quires the output of a pre-trained AE/GE Model. The STS Model requires all previous models (AE/GE, SE, STP) for its training. Finally, training the Vocoder Model requires the output of the STS Model.

3.3.1 Accent and Gender Embedding model (AE/GE Model)

To obtain fixed-length vectors representing the accent and gender properties of the speaker, the model was first trained to solve a classification task. In this configuration, class labels are used during training, whichthe model outputs at the last layer. The vector representations, used as vocal characteristics, are taken from a specific intermediate layer.

This and other models use the same Pre-processor. It is based on the Fast Fourier Transform (FFT), which converts the incoming signal from time domain into a frequency domain mel-spectrogram. This shows the frequency content of the audio signal over time on a perceptual mel scale, which approximates the non- linear frequency response of the human ear. The sampling rate is 22050 Hz, the window size is 1024 samples, the window hop size is 256 samples, and the number of generated mel bands is 80.

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI:10.15680/IJIRCCE.2025.1304229

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Figure 2 shows the training scheme of the AE/GE Model. It includes Jasper blocks configured as 3x3. Both the Accent Decoder and Gender Decoder have the same architecture, consisting of an Attention pooling layer, a normalization layer, a convolutional layer to obtain 192-dimensional vector representations (Accent Embedding, Gender Embed- ding), and a linear layer to predict the Accent Class or Gender Class.

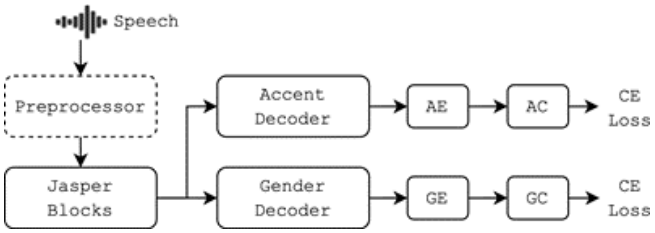


Fig. 2. Training scheme of the AE/GE Model.

3.3.2 Speaker Embedding model (SE Model)

Figure 3 shows the training scheme of the SE Model. It includes an input convolutional neural network based on the SincNet architecture, layers from the X-Vectors DNN model, and a layer for obtaining 512- dimensional vector representations.

Unlike the AE/GE Model, no preliminary conversion to a mel-spectrogram is performed. Instead, the time-domain audio signal with a sampling rate of 16000 Hz is fed into the band- pass filters of the SincNet architecture, followed by the convolutional layers of the X-Vectors DNN, and the output fully connected layer. During model training the Additive Angular Margin (AAM) Loss is minimized.

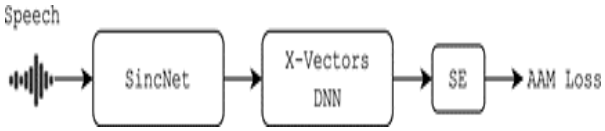


Fig. 3. Training scheme of the SE Model.

3.3.3 Speech-to-Phonemes model (STP Model)

The next step is speech recognition considering the speaker's accent. For this, a model that converts speech into phonetic or textual tokens is required. The training scheme of the STP Model is shown in Figure 4. In this figure, the blocks marked with dashed lines are fixed (or frozen) during backpropagation, meaning their weights are not updated.

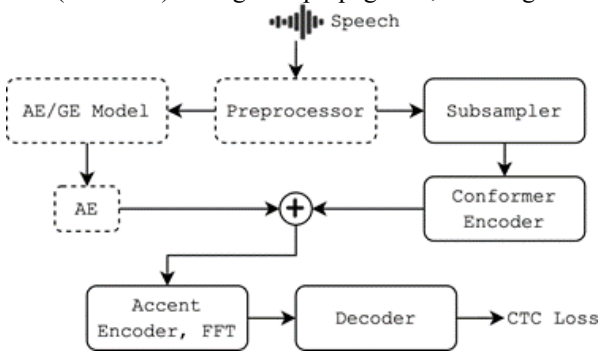


Fig. 4. Training scheme of the STP Model.

The incoming speech is fed into the Pre-processor, described earlier. Then, it is processed in parallel by the AE/GE Model to obtain the Accent Embedding (AE) and by the Sub- sampler block, which reduces the dimensionality by a factor of 4. It is then transformed in the Conformer Encoder block, which consists of 12 Conformer modules with an internal dimensionality of 512, including fully connected, convolutional, and transformer layers. Next, the accent vector is normalized, adjusted to a dimensionality of 512, summed with the output of the Conformer Encoder, and fed into the Accent Encoder block. The Accent Encoder has a Feed-Forward Transformer (FFT) architecture. The output from the

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI:10.15680/IJIRCCE.2025.1304229

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Accent Encoder is used in the STS Model as the distribution of phonetic tokens. Finally, the output from the Accent Encoder is fed intoa Decoder with a single-layer convolutional architecture and a Softmax activation function, producing a vector of predicted text tokens of a size equal to the tokenizer vocabulary (128) plus one (for the blank token). During model training, the Connectionist Temporal Classification (CTC) Loss function is minimized, which calculates the loss between the continuous (unsegmented) time series and the target sequence:

$$LSTP\ (x, y) = -\log\ (\sum \rho \in A\ \prod T\ x), (2)$$

where $LSTP$ is the loss function of the STP Model (CTC Loss); x represents the probabilities of text tokens predicted by the model; y – sequence of text tokens from the target text, ρ – alignment path that reduces the x predictions to the y sequence by removing all blank tokens and merging repeating tokens; Ax, y – set of all possible alignment paths; T – number of predicted tokens in x; $x\rho t$ – probability of a specific predicted token at step t for the selected alignment path ρ.

3.3.4 Spectrogram generation Speech-to-Speech model (STS Model)

The previous are combined into a single architecture for speech-to-speech conversion and spectrogram generation. The training scheme of the STS Model includes the previously discussed Pre-processor, STP Model, as well as the AE/GE Model and SE Model blocks with their respective vector representation modules (AE, GE, SE). All these blocks are marked with dashed lines, indicating they were pre-trained and are not updated during the STS Model training. Additionally, a non-trainable block based on the Normalized Cross-Correlation Function and median smoothing is included for extracting the fundamental frequency (F0), which is the lowest frequency of a periodic sound signal perceived as pitch by the human ear.

The incoming audio signal is processed by the Preprocessor, Pitch block, and SE Model. The mel- spectrogram from the Pre-processor is fed into the AE/GE Model and STP Model. The phonetic token distribution from the STP Model goes to the Upsampler block with a factor of 4 to align the original and generated spectrograms. The Upsampler consists of two 1D transposed convolutional layers and two ReLU activation functions, each placed after a convolutional layer. After the Upsampler, the phonetic representations are transformed by the STP Encoder, which has an architecture of six feed-forward Transformer (FFT) stacks used in the Fastpitch architecture as an input block operating in the token domain, with internal and external dimensionalities of 1536 and 384, respectively.

The accent, gender, speaker timbre, and pitch profile vectors are normalized and adjusted to a dimensionality of 384. The accent vector and the output from the STP Encoder are summed and fed into the Accent Encoder (one FFT stack). Similarly, the pitch, timbre, and gender vectors are summed and fed into the Speaker Encoder (one FFT stack). Thus, the Speaker Encoder aggregates voice characteristics unrelated to the accent, while the Accent Encoder determines phonetic pronunciation based on the accent. The sum of the Speaker Encoder and Accent Encoder vectors is then fed into the STS Decoder, which has six FFT stacks of the Fastpitch architecture operating in the output mel domain. Finally, the vector is projected to a dimensionality of 80 to match the original number of mel bands. During training, the loss function based on the mean squared error is minimized:

$$L_{STS}\ (x, y) = \frac{1}{\sum_{i=1}^{N} di} \sum_{i=1}^{N} d\ (y -)2, \quad (3)$$

where $LSTS$ – STS Model loss function (Mel Loss); N – number of elements (frames) in mel-spectrogram; x – predicted mel-spectrogram; y – ground-truth target mel-spectrogram; d – mel-spectrogram duration mask, used to batch the data into a fixed size, consists of values 1 ("element should be considered") and 0 ("element should not be considered"), derived from the predicted spectrogram's duration.

3.3.5 Simplified accent conversion model (Ablation Model)

For the purpose of conducting comparative experiments, a simplified version of the accent conversion model was also developed. The scheme of this model is shown in Figure 5.

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI:10.15680/IJIRCCE.2025.1304229

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

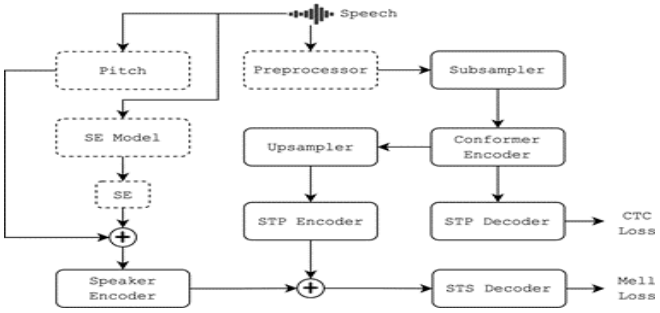


Fig. 6. Training scheme of the Ablation Model.

In this model, the vector representation module for accent and gender (AE/GE Model) and all associated encoders in the STP Model and STS Model are excluded. Thus, in the resulting Ablation Model, the output is not influenced by accent and gender properties. Additionally, the STP Model was not trained separately but simultaneously with the STS Model, without fixing the weights of the STP Model, minimizing the sum of the CTC Loss and Mel Loss functions.

IV. SUMMARY AND CONCLUSION

The non-autoregressive model proposed by Nechaev and Kosyakov outperforms previous accent conversion systems by reducing processing time and maintaining high speech quality. Their system achieves the following:

1. Low-latency accent conversion suitable for real-time applications.
2. Speaker identity preservation even after accent modification.
3. Scalability across a wide range of accents without the need for parallel data.

The introduction of TTS-guided systems has further enhanced the scalability of accent conversion models by enabling real-time accent translation across a variety of applications, including customer service and virtual assistants.The literature on accent conversion has evolved significantly, with recent advancements in non- autoregressive neural networks offering real-time, scalable solutions to accent translation. The Nechaev and Kosyakov model represents a breakthrough in low-latency speech processing, enabling real-time accent conversion while preserving speaker identity.Future work should focus on improving generalization for lesser-known accents and refining zero-shot learning techniques to further reduce data requirements. As accent conversion technology continues to improve, its applications in global communication, customer service, and language learning will become increasingly important.

REFERENCES

1. Nechaev, V., & Kosyakov, S. (2024). Non-autoregressive Real-time Accent Conversion model with voice cloning. Ivanovo State Power Engineering University. arXiv preprint arXiv:2405.13162.
2. Birner B. Why Do Some People Have an Accent? – Linguistic Society of America, Washington, DC., 1999.
3. Baese-Berk M. M., Morrill T. H. Speaking rate consistency in native and non-native speakers of English
4. .The Journal of the Acoustical Society of America. –2015. – T. 138. – №. 3. – C. EL223-EL228.
5. Piske T., MacKay I. R. A., Flege J. E. Factors affecting degree of foreign accent in an L2: A review Journal of phonetics. – 2001. – T. 29. – №. 2. – C. 191-215.
6. Munro M. J., Derwing T. M. Foreign accent, comprehensibility, and intelligibility in the speech of second language learners // Language learning. – 1995. – T. 45. – №. 1. – C. 73-97.
7. Lev-Ari S., Keysar B. Why don't we believe non-native speakers? The influence of accent on credibility
8. // Journal of experimental social psychology. – 2010. – T. 46. –№. 6. – C. 1093-1096.
9. Muniraju Hullurappa, Mohanarajesh Kommineni, “Integrating Blue-Green Infrastructure Into Urban Development: A Data-Driven Approach Using AI-Enhanced ETL Systems,” in Integrating Blue-Green Infrastructure Into Urban Development, IGI Global, USA, pp. 373-396, 2025.
10. Rubin D. L., Smith K. A. Effects of accent, ethnicity, and lecture topic on undergraduates' perceptions of non-native English-speaking teaching assistants Intenational journal of intercultural relations. – 1990. – T. 14. – №. 3. – C. 337-353.
11. Nelson Jr L. R., Signorella M. L., Botti K. G. Accent, gender, and perceived competence Hispanic Journal of Behavioral Sciences. – 2016. – T. 38. – №. 2. – C. 166-185.
12. Pinget, A. F., Bosker, H. R., Quené, H., De Jong, N. H.Native speakers’ perceptions of fluency and accent in L2 speech Language Testing. – 2014. – T. 31. – №. 3. – C. 349-365.
13. Barkhudarova E. Methodological problems in analyzing foreign accents in Russian speech Vestnik Moskovskogo universiteta. Seriya 9. Filologiya. – 2012. – №. 6. – C. 57-70.
14. Felps D., Bortfeld H., Gutierrez-Osuna R. Foreign accent conversion in computer assisted pronunciation training Speech communication. – 2009. – T. 51. – №. 10. – C. 920-932.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details