Research Article

Subhayu Ghosh, Nanda Dulal Jana, Tapas Si, Saurav Mallik*, and Mohd Asif Shah*

# CCLCap-AE-AVSS: Cycle consistency loss based capsule autoencoders for audio–visual speech synthesis

https://doi.org/10.1515/jisys-2023-0171
received September 17, 2023; accepted February 15, 2024

**Abstract:** Audio–visual speech synthesis (AVSS) is a rapidly growing field in the paradigm of audio–visual learning, involving the conversion of one person's speech into the audio–visual stream of another while preserving the speech content. AVSS comprises two primary components: voice conversion (VC), which alters the vocal characteristics from the source speaker to the target speaker, followed by audio–visual synthesis, which creates the audio–visual presentation of the converted VC output for the target speaker. Despite the progress in deep learning (DL) technologies, DL models in AVSS have received limited attention in existing literature. Therefore, this article presents a novel approach for AVSS utilizing capsule network (Caps-Net)-based autoencoders, with the incorporation of cycle consistency loss. Caps-Net addresses translation invariance issues in convolutional neural network approaches for effective feature capture. Additionally, the inclusion of cycle consistency loss ensures the retention of content information from the source speaker. The proposed approach is referred to as cycle consistency loss-based capsule autoencoders for audio–visual speech synthesis (CCLCap-AE-AVSS). The proposed CCLCap-AE-AVSS is trained and tested using VoxCeleb2 and LRS3-TED datasets. The subjective and objective assessments of the generated samples demonstrate the superior performance of the proposed work compared to the current state-of-the-art models.

**Keywords:** voice conversion, audio–visual synthesis, autoencoder, capsule network, cycle consistency loss

## 1 Introduction

Audio–visual speech synthesis (AVSS) is an exciting frontier within the realm of artificial intelligence (AI) research, holding significant promise and potential in various applications, particularly within the domain of audio–visual learning. Within the traditional framework of voice conversion (VC), the process entails converting the speech produced by a source speaker into the distinct speech characteristics of a designated target

*** Corresponding author: Saurav Mallik**, Department of Environmental Health, Harvard T H Chan School of Public Health, Boston, MA 02115, United States of America, e-mail: sauravmtech2@gmail.com, smallik@hsph.harvard.edu
*** Corresponding author: Mohd Asif Shah**, Department of Economics, Kebri Dehar University, Jigjiga 3060, Ethiopia; Centre of Research Impact and Outcome, Chitkara University Institute of Engineering and Technology, Chitkara University, Rajpura, 140401, Punjab, India; Chitkara Centre for Research and Development, Chitkara University, Baddi, 174103, Himachal Pradesh, India,
e-mail: drmohdasifshah@kdu.edu.et
**Subhayu Ghosh:** Department of Computer Science and Engineering, National Institute of Technology Durgapur, West Bengal 713209, India, e-mail: subhayu.ghosh17@gmail.com, sg.22cs1101@phd.nitdgp.ac.in
**Nanda Dulal Jana:** Department of Computer Science and Engineering, National Institute of Technology Durgapur, West Bengal 713209, India, e-mail: ndjana.cse@nitdgp.ac.in
**Tapas Si:** Department of Computer Science and Engineering, AI Innovation Lab, University of Engineering and Management, Jaipur, Rajasthan 303807, India, e-mail: c2.tapas@gmail.com

speaker, all while retaining the underlying linguistic content of the utterance [1–4]. However, recent advancements have taken this concept further by incorporating visual cues into the VC process, showcasing the potential to enhance the clarity and understandability of the synthesized speech significantly. By fusing visual information with auditory cues, this incorporation has the potential to fortify the resilience of the VC process [5]. Remarkably, this development has far-reaching implications, finding applications across various domains, including multimedia, education, animation, and even forensic studies [6,7]. Moreover, the essence of this technique lies in the intricate procedure of seamlessly transfiguring the voice characteristics and the accompanying visual attributes of the source speaker into the unified audio–visual manifestation of the chosen target speaker. Throughout this process, the primary objective remains preserving the essential speech content, ensuring a coherent and faithful transformation. Consequently, this synthesis of auditory and visual information emerges as a distinct field recognized and defined as AVSS.

This intricate process involves two critical sequential components: the VC stage, which is responsible for translating vocal characteristics from the source speaker into those resembling the target speaker, and the audio–visual synthesis (AVS) stage, which is tasked with creating the audio–visual output of the transformed speech produced by VC, tailored to the desired target speaker. An overview of the AVSS framework is presented in Figure 1, illustrating the step-by-step procedure by which vocal traits from both the source and target speakers are initially extracted. Subsequently, the features of the source speaker undergo a transformation process to replicate the target speaker's voice during the reconstruction phase, as depicted in the VC model block. Following this, the converted audio attributes of the target speaker traverse through the pre-trained AVS model, which has previously executed the visual alignment of the target speaker's original audio. This process culminates in creating an audio–visual stream from the converted speech of the target speaker, as illustrated in the AVS block.

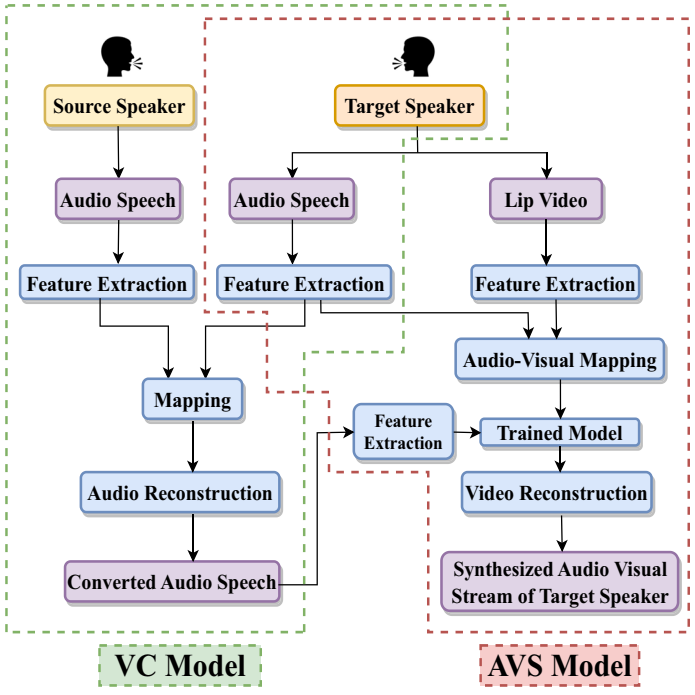


**Figure 1:** Overview of the AVSS process.

It is noteworthy that the development of deep learning (DL) models for AVSS has been relatively limited, especially when compared to other research domains. As far as our knowledge extends, the existing literature has exclusively utilized Autoencoder (AE)-based generative models for AVSS [8]. An AE is a feed-forward

neural network model consisting of convolutional neural network (CNN) building blocks [9,10]. However, CNNs encounter challenges when it comes to effectively categorizing objects with diverse orientations due to their inherent property of translation invariance. It becomes particularly evident in the context of AVSS, where the arrangement of speech features in the latent space can significantly influence their pronunciation. Consequently, the traditional CNN architecture proves to be inadequate for capturing these nuanced variations. Hence, to surmount this limitation and successfully capture the intricate variations essential for AVSS, there arises a compelling need for an advanced neural network model that can adeptly accommodate and comprehend these complexities.

The capsule network (Caps-Net) stands out as a distinctive neural network model characterized by its composition of capsules. These capsules have garnered attention recently as a promising alternative to CNNs specifically for tasks involving classification [11–13]. Unlike conventional CNNs, the Caps-Net leverages dynamic routing algorithms to facilitate intricate communication and harmonization among capsules across distinct layers. This strategic interaction aims to transmit precise information regarding their individual features, thus culminating in the generation of more intricate and comprehensive representations. This intrinsic attribute holds the potential to render the Caps-Net exceptionally well suited for AVSS. Moreover, it is imperative to ensure the integrity of contextual information embedded within speech content throughout the VC process. A potential avenue for achieving this objective lies in the integration of a cycle consistency loss mechanism, a measure that has demonstrated the capability to enhance the robustness and adaptability of the AVSS procedure. By upholding the fidelity of translation while adhering to cyclical consistency principles [14–16], this augmentation has the potential to induce a substantial elevation in the overall quality of the translation output. In the pursuit of refining the training process of the neural network, an essential consideration revolves around enhancing the adaptive learning rate for expediting convergence. It is where the Rectified Adam (RAdam) optimizer [17] steps in, addressing the pertinent challenge by imbuing a stabilizing influence on the convergence trajectory.

In this research endeavor, our contributions encompass two distinct models to enhance the efficiency of the AVSS framework significantly: first, a Caps-Net Adversarial Autoencoder (AAE) tailored to VC, and second, a Caps-Net AE designed for AVS. Specifically, the Caps-Net architecture assumes a pivotal role within the discriminator of our AAE model, strategically crafted for VC. Simultaneously, this Caps-Net architecture is responsible for serving as the encoder within the AE model, meticulously crafted to generate an audio–visual stream of the target speaker. To further amplify the faithful retention of the inherent content information from the source speaker, we ingeniously incorporate the cycle consistency loss into the fabric of the AAE model. Moreover, to expedite the convergence of the training process and enhance optimization, we have adopted the RAdam optimizer. The culmination of these components leads us to christen our AVSS framework as “cycle consistency loss-based capsule autoencoders for AVSS,” succinctly termed as CCLCap-AE-AVSS.

We extensively train and test the proposed model on two significant datasets: VoxCeleb2 [18] and LRS3-TED [19]. Furthermore, for the purpose of validating the efficacy of our VC model, we have also leveraged the VCTK dataset [20]. To gauge the effectiveness and performance of CCLCap-AE-AVSS, the comprehensive evaluation encompasses both objective metrics and subjective human assessments of the synthesized samples. The conclusive results of these multifaceted evaluations collectively and unequivocally establish the unparalleled superiority of our proposed approach over existing state-of-the-art (SOTA) models. This outcome underscores the potential of our proposed CCLCap-AE-AVSS framework to advance the realm of AVSS in terms of both audio and video quality.

In summary, this work has been enriched by several notable and impactful contributions, which can be outlined as follows:

- We propose a novel Caps-Net discriminator within the AAE framework for the VC model, aiming to efficiently distinguish speech features.
- We have integrated Caps-Net in the encoder of the AVS model to improve the quality of the generated video.
- The VC model incorporates cycle consistency loss to preserve the content information of the source speaker.
- We utilize RAdam optimizer to accelerate convergence speed, leading to an enhancement in model accuracy.

- It has been observed that CCLCap-AE-AVSS has demonstrated a significant improvement in audio and video quality compared to the SOTA models.

The rest of this article is organized as follows: the essential fundamental preliminaries of the proposed framework are presented in Section 2. Section 3 discusses about the prior relevant works on AVSS and its components, while the proposed approach is elaborated in Section 4. Section 5 includes the dataset description, feature details, network architecture of CCCLap-AE-AVSS, training details, considered SOTA models, and system configuration. The detailed experimental results and the ablation studies of the proposed model are presented in Section 6. In Section 7, we present the limitations and disadvantages of the proposed method, and also the possible future research scopes to overcome the problems. Finally, Section 8 concludes this article.

# 2 Background

In this section, we provide a concise yet comprehensive overview of the essential foundations that underlie our proposed framework. The fundamental prerequisites and initial building blocks of our conceptual framework are expounded upon and explored in a succinct manner within this segment.

## 2.1 AAE

An AAE refers to a special type of neural network model that integrates the concepts of AE and adversarial training methodologies [21]. The AAE structure consists of two parts: first, an encoder–decoder pair and followed by a discriminator. The role of the encoder is to transform input data into a latent representation having lower dimensions [9,22], while the decoder's task is to reconstruct the input data from this latent representation. In addition to these components, a discriminator, also referred to as a critic, is incorporated to differentiate between real latent representations derived from the input data and fake latent representations sampled from a prior distribution. The primary objective of the discriminator [23,24] is to distinguish between these genuine and synthetic latent representations. The mathematical formulations that describe the functionality of the encoder, decoder, and discriminator are presented in equations (1)–(3), respectively.

$$\text{Encoder} : A \rightarrow X, \tag{1}$$

$$\text{Decoder} : X \rightarrow \hat{A}, \tag{2}$$

$$\text{Discriminator} : X \rightarrow \text{real/fake}, \tag{3}$$

where $A$ is input data, $X$ is the latent representation, and $\hat{A}$ is the reconstructed output by the decoder of the AE.

Adversarial training involves refining the encoder with the goal of generating latent representations that are difficult for the discriminator to differentiate from genuine examples. This complex process drives the encoder to create latent representations that are rich in meaning and carry significant information. Through this method, the encoder is continuously challenged and improved, leading to the production of high-quality, informative latent representations that closely resemble authentic data. The adversarial dynamic between the encoder and the discriminator ensures that the encoder's output becomes increasingly indistinguishable from real data, thereby enhancing the overall quality and utility of the representations.

## 2.2 Caps-Net

Caps-Net represents a recently developed model in deep neural networks. It consists of convolutional, primary, and digit caps layers. The role of the convolutional layer is to extract high-level feature vectors from the

input data. These extracted feature vectors maintain their original information while reshaping within the primary caps layer. In this layer, the activation function employs the squashing algorithm. This algorithm probabilistically determines the pre-existing feature vectors' orientation and relative spatial relationships. Subsequently, the feature vectors undergo dynamic routing [25] toward the digit caps layer. The ultimate purpose of this dynamic routing is to prepare the feature vectors for the final classification stage. It is noteworthy that Caps-Net has shown its potential ability to identify the orientation changes of features by monitoring the overall information shift or the feature correlation within the input [26]. Moreover, it requires less training data to generalize well compared to CNN models, leading to improved performance in various tasks more effectively [27]. The architecture of Caps-Net is depicted in Figure 2.

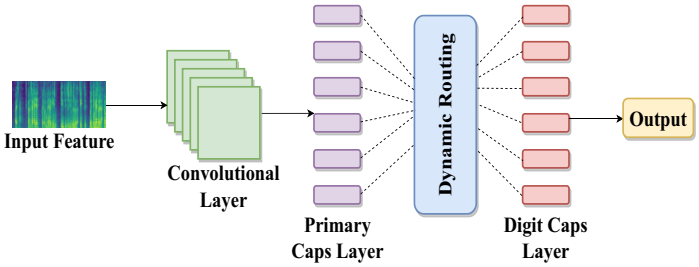


**Figure 2:** Caps-Net architecture.

### 2.3 RAdam optimizer

RAdam optimizer [17] is an optimization algorithm designed to improve upon the popular Adam optimizer by addressing some of its limitations. RAdam adjusts the adaptive learning rate components of Adam using a variance correction term, which helps stabilize training and convergence. The variance correction term is computed to address the biased estimates of the first and second moments of the gradients. The corrected moving average of squared gradients is given by

$$\tilde{v}_t = \frac{v_t}{1 - \beta_2^t}, \tag{4}$$

where $v_t$ is the moving average of squared gradients at step $t$, and $\beta_2^t$ is the decay rate for the squared gradients.

## 3 Related work

We are presenting a novel framework dedicated to AVSS, offering a heightened level of robustness compared to the conventional VC technique. Therefore, in this section, our focus is directed toward an exploration of the existing body of research pertaining to VC and the burgeoning field of AVSS.

### 3.1 VC

VC constitutes a prominent and dynamic domain within the realm of speech processing. Its primary objective is the manipulation of the distinct vocal attributes of a speaker, all the while upholding the essential linguistic essence and preserving the inherent natural quality. With the progressive development of research, numerous methodologies and strategies have been applied with the goal of generating synthetic speech of the utmost caliber [3,28]. In the primary stages of VC exploration, it relied upon the realm of statistical methodologies. This included the employment of tools such as Gaussian mixture models [29] and hidden Markov models [30]. These

pioneering models paved the way for the transformation of spectral attributes across different speakers. However, these models often grappled with the challenge of faithfully retaining the intricate rhythms of prosody and the innate authenticity that characterizes human speech [31].

With the emergence of DL technologies, the field of VC has undergone significant advancement. A particularly noteworthy milestone in this trajectory was the CycleGAN-based VC framework by Kaneko and Kameoka [32]. This innovative approach harnessed the power of adversarial training, with the transformation of source features into target features without requiring parallel datasets. The outcomes achieved by this methodology were remarkable, as evidenced by the substantial improvements in voice quality and the reduced reliance on parallel training data. However, this framework relied on paired datasets, where corresponding source and target speakers were aligned. Handling non-parallel data, where there was no one-to-one correspondence between source and target utterances, was a challenge with this model. Moreover, the CycleGAN-VC model was highly dependent on the quality and diversity of the training data. Furthermore, the realm of VC has continued to evolve, with recent research delving into the realms of cross-lingual and multi-lingual VC [33]. These novel approaches seek to transcend language barriers and speaker diversity by enabling the conversion of speech between disparate languages or among speakers with distinct linguistic backgrounds. The cross-lingual VC technique functions by aligning phonetic contexts between the source and target languages, thereby showcasing the immense potential of VC in a wide array of linguistic scenarios.

However, the VC framework can be more robust and improved with the integration of visual cues. This integration engenders a more harmonious and coherent representation of a speaker and also augments the overall robustness of the VC process. By fusing auditory and visual information, the VC paradigm gains the ability to generate outputs that are consistently faithful to the speaker's identity, thus pushing the boundaries of its capabilities.

## 3.2 AVSS

AVSS involves creating synchronized speech and corresponding facial movements. This technology has attracted considerable interest because of its potential uses in areas such as entertainment, communication aids for the hearing impaired, and virtual human interactions. Initial attempts in AVSS predominantly employed rule-based and template-based techniques. A notable early contribution by Cassell et al. [34] presented a rule-based system that produced lip movements from phoneme sequences. Nevertheless, these early AVSS methods were often limited in their expressiveness and faced challenges in replicating natural speech dynamics.

Subsequently, the amalgamation of statistical models as well as machine learning (ML) techniques brought significant advancements to AVSS. Barbulescu et al. pioneered joint conversion methodologies for audio and visual features, integrating prosodic elements into their model [5]. They improved the VC method by incorporating 3D facial expressions as the visual data. A comprehensive analysis of both feature sets was conducted to assess their results. However, their approach was hindered by the lack of available datasets and the limited experiments in their research.

Sawada et al. proposed audio–visual voice conversion (AVVC) technique to combine audio features with the visual data [35]. For visual inputs, they used lip images and applied principal component analysis to derive eigenlips, which served as visual parameters [36]. In terms of audio feature extraction, the source speaker's features were obtained using the fast Fourier transform, while target speakers' audio features with high resolution were extracted using the STRAIGHT method. Their objective was to improve the quality of converted speech, particularly in noisy environments. However, their method struggled to achieve perfect synchronization between audio and lip movements.

Later, Tamura et al. introduced an AVVC approach utilizing Deep BottleNeck Features (DBNF) [37] and deep canonical correlation analysis (DCCA) [38] to advance traditional methods [39]. DBNF was employed to enhance feature representations, while DCCA generated more correlated features across different modalities and refined features based on modality. They also developed a novel cross-modal VC tool that was effective for both audio and visual features using DCCA. Although this approach demonstrated potential in the field, it was limited by high computational demands and the complexity of the model required to produce synthesized outputs.

With the progression of DL, researchers have investigated innovative methods for AVSS by utilizing neural networks. These models employed CNNs to understand complex relationships between audio and visual features. A significant contribution by Deng et al. [8] introduced a novel AVSS technique known as exemplar AEs. This method enabled the conversion of speech from an unknown speaker into the unique voice of a known target speaker. This approach allows the transformation of various audio inputs into the audio–visual streams of multiple distinct target speakers, marking the beginning of many-to-many AVSS. However, this model struggled with producing natural-sounding audio and high-quality video, and it did not effectively capture input features.

Despite its promising potential, the field of AVSS using DL is less explored compared to other domains. As the field progresses, integrating more complex and advanced architectural designs will likely produce even more robust results in AVSS. The trend suggests a future where the combination of sophisticated neural network architectures will lead to significant advancements in seamless AVSS.

# 4 Proposed approach

Within this section, a comprehensive analysis is conducted to delve into the intricacies of the proposed CCLCap-AE-AVSS framework. Additionally, a thorough exploration of the workflow underpinning the proposed approach for AVSS is provided. Furthermore, we dedicate space to elucidating the assortment of loss functions that have been thoughtfully taken into account within the framework's scope, enriching our understanding of its architectural essence.

## 4.1 Model description of CCLCap-AE-AVSS

In this work, we propose two models referred to as Caps-Net AAE-based VC and Caps-Net AE-based AVS, for the purpose of AVSS. The initial model is focused on altering vocal characteristics from a source speaker to a target speaker, all while retaining the speech content. The subsequent model is trained using the original voice of the target speaker along with the corresponding audio–visual stream. It generates a video with visual mapping that relies on the transformed speech output generated by the first model. The AAE-based VC model includes encoder, decoder, and discriminator in its architecture, while the AE-based AVS model comprises encoder and decoder. The Caps-Net is utilized both in the discriminator of the VC model and in the encoder of the AVS model. Additionally, the first model also incorporates cycle consistency loss. We have named this proposed framework as CCLCap-AE-AVSS. The comprehensive architecture of both components of the CCLCap-AE-AVSS is depicted in Figure 3.

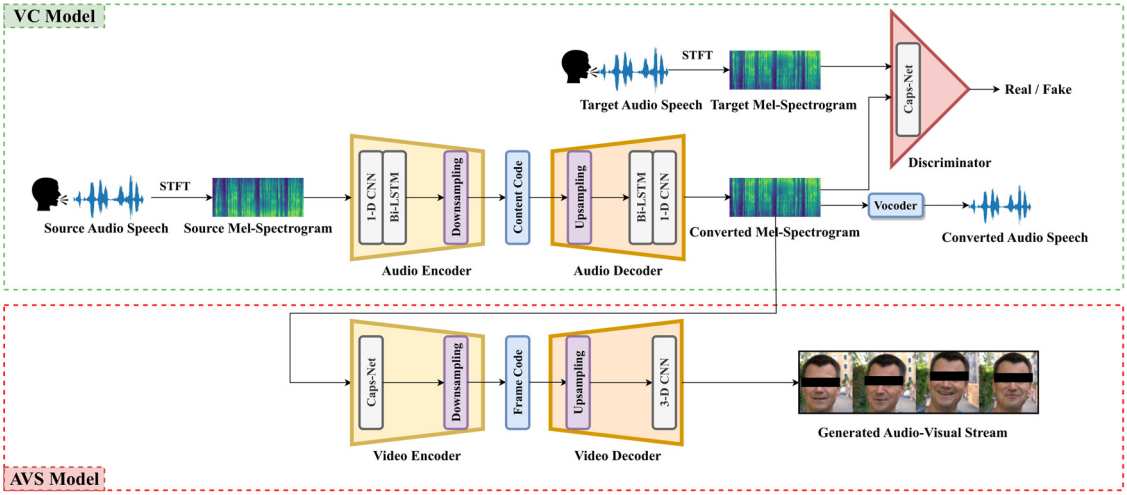


**Figure 3:** Proposed CCLCap-AE-AVSS framework.

### 4.1.1 VC model

The primary aim of the VC model is to alter the vocal characteristics of the original speaker's voice so that it closely mimics the voice of a target speaker, while maintaining the original speech content. Formant patterns in the mel-spectrogram represent both unique features of individual speakers and more universal elements. Modifying these patterns, such as by shifting their position or rearranging them, can potentially disrupt the crucial informational cues embedded in the speech features.

Caps-Net offers a robust solution to this problem due to its ability to detect changes in the orientation of formant patterns. To address this complex issue, we incorporate Caps-Net into the discriminator component of the AAE within the VC model. The discriminator acts as an essential evaluator, assessing the similarity between the genuine speech features and those generated by the VC model. By integrating Caps-Net, we leverage its strengths to improve feature extraction and effectively manage the spatial hierarchies involved in distinguishing between real and synthesized data. Thus, embedding Caps-Net in the AAE's discriminator significantly advances the fidelity and naturalness of the converted speech output.

Mathematically, the Caps-Net discriminator $D$ can be represented as

$$D(.) \rightarrow \text{CapsNet}(.), \tag{5}$$

where the output of discriminator $D$ is the last layer feature representation obtained from the digit caps layer.

We have integrated cycle consistency loss into our VC model to achieve higher levels of naturalness and robustness in the converted speech. This integration is essential for preserving the unique speech content of the source speaker while enabling cross-lingual VC capabilities. Our approach also supports bi-directional conversion, ensuring consistent transformation between the source and target speakers in both directions. In our methodology, we use mel-spectrogram features, labeled $m_X$ and $m_Y$, extracted from the audio of the source speaker $X$ and the target speaker $Y$. These features are then processed by an encoder, a crucial component of our model. The encoder individually processes the mel-spectrograms $m_X$ and $m_Y$, producing corresponding content codes $z_X$ and $z_Y$. This procedure is illustrated in Figure 4.

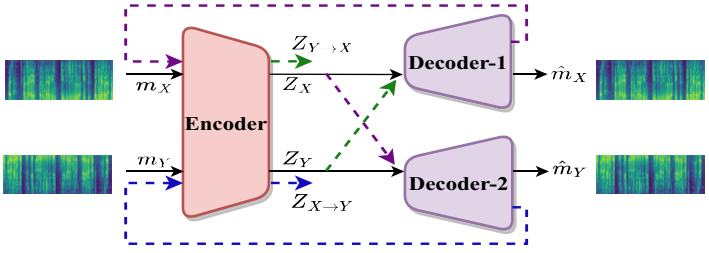


**Figure 4:** Computation of cycle consistency loss.

The decoders subsequently produce the transformed mel-spectrograms, labeled as $\hat{m}_X$ and $\hat{m}_Y$ for their respective domains. Next, the content codes, which carry the core information of the input, are sent to the decoders of the opposite domains. This cross-domain processing creates a disparity between the nature of the content codes and the function of the decoders, resulting in a noticeable mismatch. To correct this, the mismatched content codes are cycled back to the encoder to compute the cycle consistency loss. This metric measures the degree of mismatch that has occurred. The main goal is to assess how well the content codes and decoders align during this transformative cycle. Integrating this mechanism into the model ensures efficient retention of content-related information throughout the VC process.

#### 4.1.2 AVS model

The AVS model's primary goal is to generate a synchronized audio–visual stream tailored to a designated target speaker. This synchronization is achieved by transforming the source speaker's speech through the AVS model. Initially, the AVS model undergoes training using genuine audio recordings and corresponding visual data from the target speaker, establishing the model's understanding of the target speaker's unique characteristics.

Following this, the output from the VC model is used as input for the AVS model. Our proposed CCLCap-AE-AVSS approach integrates Caps-Net into the encoder section of the AE associated with the AVS model. This integration offers significant benefits, notably reducing the need for numerous filters typically required to detect various feature manifestations. As a result, the model operates more efficiently in terms of parameter utilization compared to traditional CNNs. Additionally, this approach lowers the data requirements for model training due to Caps-Net's inherent ability to capture pose information, a fundamental aspect of its architecture. Furthermore, Caps-Net demonstrates a strong capacity for generalization when encountering new data, enhancing the AVS model's overall performance and enabling it to generate high-quality videos.

## 4.2 Working principle of CCLCap-AE-AVSS

The step-by-step workflow of the proposed framework is illustrated in Algorithm 1. This comprehensive process involves employing the VC and AVS models to facilitate the entire AVSS procedure. To commence, we initiate the sequence of actions by capturing the audio speech of the source speaker, which is intended for transformation into the voice of the target speaker. Furthermore, a sample of audio–visual speech is obtained from the target speaker, from which the corresponding audio speech is extracted [40]. The audio attributed to the source speaker is denoted as $X$, while the audio–visual sample of the target speaker is labeled as $AV_Y$ and the extracted audio of the target speaker is referred to as $Y$.

**Algorithm 1: Workflow of CCLCap-AE-AVSS**

**Initialize:** $X$, $AV_Y$.
$Y = extract_{audio}\ (AV_Y)$.
$m_X = STFT(X)$,
$m_Y = STFT(Y)$.
**while** $Train_{\mathrm{V\,C}}\ (m_X, m_Y)$: **do**
  $C_X = (Input_{AEnc}(m_X))$.
  **while** $(downsample(C_X))$ **do**
    $Input_{ADec} = (downsample(C_X))$,
    $Output_{ADec} = (Upsample(Input_{ADec}))$,
    $m_Z = Output_{ADec}$.
  **end**
  $Z = Wavenet(m_Z)$.
**end**
**while** $Train_{\mathrm{AV\,S}}(Y, AV_Y)$: **do**
  $Input_{\mathrm{AV\,S}} = m_Z$
  for each frame in $AVS$ : align-lip-movements (frame,$m_Z$).
  $Output_{\mathrm{AV\,S}} = AVS(\mathrm{Input}_{\mathrm{AV\,S}})$.
**end**

Following this, both sets of speech undergo a short-time Fourier transform (STFT) process [41], resulting in their conversion into mel-spectrogram representations with dimensions of 80 × 128. These transformed mel-spectrograms are specifically referred to as $m_X$ for the source speaker and $m_Y$ for the target speaker. Subsequently, these mel-spectrograms are inputted into the audio encoder, which is referred to as *AEnc*, yielding the respective content codes $C_X$ and $C_Y$. The content code of the source speaker is then transmitted to the audio decoder defined as *ADec* after undergoing downsampling. The decoder undertakes the process of upsampling, culminating in the creation of the converted mel-spectrogram, denoted as $m_Z$. In order to transcribe this transformed mel-spectrogram into the speech signals of the target speaker, Wavenet vocoder is employed. Here, the converted audio of the target speaker is denoted as $Z$.

Moving forward, the mel-spectrogram of the converted audio speech from the target speaker is fed into the AVS model. This particular AVS model has previously undergone training utilizing audio–visual data from the target speaker, effectively incorporating visual cues to accurately represent their original voice. Through meticulous adjustments carried out on a frame-by-frame basis, aligning the lip movements with the corresponding audio speech, we are able to successfully synthesize the audio–visual stream of the converted speech.

## 4.3 Loss functions

### 4.3.1 Cycle consistency loss

Our proposed CCLCap-AE-AVSS incorporates cycle consistency loss in its VC model. As indicated in Figure 4, this loss can be computed as

$$\mathcal{L}_{\text{Cycle}} = ||z_{X\to Y} - z_X||^2 + ||z_{Y\to X} - z_Y||^2, \tag{6}$$

where $\mathcal{L}_{\text{Cycle}}$ is the cycle consistency loss of the VC model and a type of $L_1$ loss. Here, $z_{X\to Y}$ and $z_{Y\to X}$ refers to the mismatched content code from encoder to decoder-1 to decoder-2, respectively.

### 4.3.2 Audio reconstruction loss

The process of converting $m_X$ and $m_Y$ into their respective reconstructed forms, $\hat{m}_X$ and $\hat{m}_Y$, takes into account an additional factor known as the audio reconstruction loss. This particular loss component plays a crucial role in the VC process. This loss factor serves as a pivotal aspect of the overall transformation process, contributing to the fidelity and accuracy of the model. This loss can be calculated as

$$\mathcal{L}_{\text{Audio Recon}} = ||\hat{m}_X - m_X||^2 + ||\hat{m}_Y - m_Y||^2. \tag{7}$$

### 4.3.3 Vocoder loss

In the VC framework, employing a vocoder is crucial. This vocoder is instrumental in producing the output speech for the target speaker from the mel-spectrogram, which is derived from the speech of the source speaker. A key aspect of the vocoder's training process involves deriving a specialized loss function, known as cross-entropy loss. This loss function is vital for optimizing the vocoder's parameters. By incorporating this loss, denoted as $\mathcal{L}_{\text{Vocoder}}$, the vocoder adjusts its internal parameters to accurately convert mel-spectrograms into coherent speech signals.

### 4.3.4 Total loss for VC

Subsequently, the cumulative loss of the VC model will be calculated as the aggregate of three distinct loss functions. It can be represented as:

$$\mathcal{L}_{\text{VC}} = \mathcal{L}_{\text{Cycle}} + \mathcal{L}_{\text{Audio Recon}} + \mathcal{L}_{\text{Vocoder}}, \tag{8}$$

where $\mathcal{L}_{\text{VC}}$ refers to the total loss of the VC process.

#### 4.3.5 Video reconstruction loss

To facilitate the process of reconstructing the video of the target speaker by utilizing the converted speech uttered by the source speaker, a crucial step involves the computation of a $L_1$ loss. This particular loss can be elegantly expressed as $\mathcal{L}_{\text{Video Recon}}$, wherein the symbols $v$ and $\hat{v}$ hold the significant roles of representing the unaltered original video and the video that has been meticulously reconstructed through the decoder, respectively:

$$\mathcal{L}_{\text{Video Recon}} = ||v - \hat{v}||_1. \tag{9}$$

#### 4.3.6 Total loss for AVS

Since the AVS model is intricately reliant upon the outputs generated by the VC model, the cumulative loss it experiences becomes inherently linked to the losses incurred by the VC model itself. In essence, the total loss encountered by the AVS model is an amalgamation of two crucial components: the comprehensive VC loss and the loss associated with the reconstruction of the video output. Consequently, the complete loss function for the AVS model can be expressed as

$$\mathcal{L}_{\text{AVS}} = \mathcal{L}_{\text{VC}} + \mathcal{L}_{\text{Video Recon}}, \tag{10}$$

where the symbol $\mathcal{L}_{\text{AVS}}$ serves as a representation of the aggregate loss sustained by the AVS model due to these intertwined factors.

# 5 Experimental setups

## 5.1 Dataset description

Our proposed model, CCLCap-AE-AVSS, has been rigorously trained and tested using the VoxCeleb2 [18] and LRS3-TED [19] datasets to thoroughly evaluate its performance. The VoxCeleb2 dataset contains 150,480 audio–visual speech segments from 6,112 unique individuals. Notably, approximately 61% of the speakers in this dataset are male. It is organized into distinct subsets for training, testing, and validation, ensuring that models are trained and evaluated on different speakers and recordings, which aids in assessing their generalization capabilities.

The LRS3-TED dataset includes video content of 400 h, meticulously gathered from 5,594 TED and TEDx talks in English, sourced from YouTube, having a wide range of speakers. The face tracks are available as mp4 files with 224 × 224 pixel resolutions and play at 25 fps, while the audio tracks are in a 16-bit format with a 16 kHz sampling rate. The dataset is divided into three sets: pre-train, train-val, and test, with the test set being completely independent from the others.

Additionally, the proposed VC model has been validated using the VCTK dataset [20], which includes 109 speakers, offering a diverse array of genders, accents, and ages. These audio recordings were captured in a controlled environment to reduce the background noise, ensuring high-quality speech content. The sound-tracks are typically available in WAV format with 48 kHz sampling rate.

## 5.2 Feature details

We choose mel-spectrogram [42] as speech feature in the proposed CCLCap-AE-AVSS framework due to its effectiveness in capturing important characteristics of speech signals. Mel-spectrograms focus on the acoustic characteristics of speech while being relatively invariant to linguistic content. This is important for VC, where the goal is to modify the speaker identity while preserving the linguistic information and prosody of the speech.

In our AVSS framework, we incorporate individual frame-by-frame lip images [43] as the visual feature. By integrating these finely detailed lip images into CCLCap-AE-AVSS, we aim to achieve a more accurate and natural synchronization between the spoken words and the corresponding lip movements, resulting in more realistic and effective AVSS system.

## 5.3 Network architecture of CCLCap-AE-AVSS

CCLCap-AE-AVSS framework[1] presents a comprehensive amalgamation of sophisticated components, ranging from audio encoder, decoder, and discriminator within the VC model to video encoder and decoder within the AVS model. These components synergistically contribute to the framework's capability to perform intricate AVSS tasks, encapsulating many complex processing steps.

For the VC model, the audio encoder is intricately designed and is composed of three 1-D CNN layers and two bidirectional long short-term memory (LSTM) (bi-LSTM) layers. To extract relevant features, a kernel size of 5 is employed, and the activation function ReLU is utilized within the audio encoder. On the other hand, the audio decoder mirrors this intricate design by featuring three 1-D CNN layers and an equal count of three bi-LSTM layers. Just like the encoder, the decoder also adopts a kernel size of 5 and relies on the ReLU activation function for optimal performance. The discriminative element within the VC model is ingeniously fashioned using Caps-Net. This involves incorporating a 2-D convolutional layer with the ReLU activation function to facilitate the extraction of high-level features. Following this, a primary caps layer, characterized by a kernel size of 9, is introduced to further process the extracted features. Ultimately, a digit caps layer is incorporated to bring about the final classification process.

In the proposed approach, the AVS model is centered on two key components: the video encoder and the video decoder. The video encoder leverages the Caps-Net architecture, selected for its suitability in handling the complexities of the AVS model. Conversely, the video decoder takes a different approach by utilizing three layers of 3-D CNNs. These layers work together to produce the audio–visual output corresponding to the target speaker. To achieve seamless integration and optimal performance, the decoder consistently uses the ReLU activation function.

## 5.4 Training details

The training process for the VC model commenced by utilizing speeches from both speakers, effectively capturing and mapping their distinct vocal features. Similarly, for the AVS model, we utilized a combination of audio speech data and synchronized frame-by-frame lip images of the target speaker, creating a comprehensive dataset that facilitated precise mapping from spoken language to corresponding lip movements in videos.

In the experimental phase, we carefully configured hyperparameters for optimal training outcomes. This included setting the learning rate to 0.001 and selecting a batch size of 8, following recommendations from the study of Deng et al. [8]. To enhance model effectiveness, we employed the RAdam optimizer [17], which addresses the slow convergence issue of traditional Adam optimizer and stabilizes training with a better adaptive learning strategy.

1 The code execution of the CCLCap-AE-AVSS is publicly shared at: https://github.com/Subhayu-ghosh/CCLCap-AE-AVSS.

Furthermore, both the VC and AVS models underwent rigorous training for 100 epochs, with batch normalization technique applied throughout the process. This extended training period ensured that CCLCap-AE-AVSS had ample exposure to data and optimization processes, ultimately enhancing their performance and capabilities.

### 5.5 SOTA models

We thoroughly compared our newly introduced CCLCap-AE-AVSS frameworks with several SOTA models. For VC, we assessed notable models such as StarGAN-VC [44], StarGAN-VC2 [45], Blow [46], MelGAN-VC [47], FLSGAN-VC [48], and Exemplar-AE [8]. For the AVS model, our evaluation included Speech2Vid [49], LipGAN [50], and Exemplar-AE [8].

StarGAN-VC [44] blends features from cycle variational autoencoder [51] and cycleGAN [32], employing a single encoder–decoder generator network regulated by auxiliary input to achieve multiple-to-multiple mappings. Unlike CycleGAN-VC, StarGAN-VC does not require attribute information during testing, resembling CVAE-VC in this aspect. StarGAN-VC2 [45] enhances domain-specific alterations using modulation-based conditional transformation.

Blow [46] utilizes a single-scale structure with forward–backward conversion and shared speaker embeddings. MelGAN-VC [47] employs spectrograms and a generative adversarial network (GAN) for domain translation, supplemented by a siamese network. FLSGAN-VC [48] integrates a self-attention mechanism and modulation spectra distance for high speaker similarity.

Speech2Vid [49] generates talking faces by combining a target speaker's static image with their audio segment using an encoder–decoder CNN architecture. LipGAN [50] facilitates face-to-face translation of speech, maintaining realistic lip synchronization through its GAN-based framework.

Exemplar-AE [8], designed for AVSS, is a straightforward encoder–decoder model constructed using CNNs. We conducted a comprehensive comparison of our framework against Exemplar-AE for both VC and AVS models throughout our experiments, consistently utilizing it as a baseline reference point.

### 5.6 System configuration

The proposed CCLCap-AE-AVSS underwent experimental testing using a Dell precision 7820 workstation. This workstation was outfitted with an Intel Xeon Gold5215 processor clocked at 2.5 GHz. The operating system in use was Ubuntu 18.04, with a 64-bit architecture, ensuring efficient utilization of the hardware's capabilities. The graphics processing was bolstered by an Nvidia 16 GB Quadro RTX-5000 graphics card, enhancing visual processing tasks. To accommodate the computational demands, the workstation boasted a substantial 96 GB of RAM, allowing for the handling of complex operations with ease. The implementation of the model was carried out using Python 3.7.0, with the assistance of PyTorch 1.1.0.

## 6 Results and discussion

### 6.1 Objective and subjective evaluations

We conducted a comprehensive assessment of the performance of our newly developed CCLCap-AE-AVSS framework through a combination of both objective [52] and subjective [53] evaluations. The objective

**Table 1:** Performance evaluation of VC models using VoxCeleb2 dataset

| Model | MCD (↓) | MSD (↓) | MOS (↑) |
|---|---|---|---|
| StarGAN-VC [44] | 6.13 | 1.97 | 2.48 |
| StarGAN-VC2 [45] | 5.98 | 1.82 | 2.89 |
| Blow [46] | 5.89 | 1.79 | 2.92 |
| MelGAN-VC [47] | 6.06 | 1.90 | 2.81 |
| FLSGAN-VC [48] | 5.69 | 1.70 | 3.21 |
| Exemplar-AE [8] | 5.65 | 1.71 | 3.10 |
| **CCLCap-AE (Ours)** | **5.30** | **1.54** | **3.98** |

**Table 2:** Performance evaluation of VC models using LRS3-TED dataset

| Model | MCD (↓) | MSD (↓) | MOS (↑) |
|---|---|---|---|
| StarGAN-VC [44] | 6.38 | 2.04 | 2.39 |
| StarGAN-VC2 [45] | 6.11 | 1.96 | 2.48 |
| Blow [46] | 5.87 | 1.99 | 2.55 |
| MelGAN-VC [47] | 6.09 | 2.03 | 2.44 |
| FLSGAN-VC [48] | 5.85 | 1.91 | 2.70 |
| Exemplar AE [8] | 5.77 | 1.81 | 2.87 |
| **CCLCap-AE (Ours)** | **5.66** | **1.69** | **3.74** |

evaluation entails a meticulous comparison between the original samples and the generated counterparts[2], achieved mathematically using metrics such as mel-cepstral distortion (MCD) [54] and MSD [55]. On the other hand, the subjective evaluation centers on gauging the human perception of the generated audio samples, employing the mean opinion score (MOS) [56] methodology. Lower values of MCD (↓) and MSD (↓) metrics inherently indicate superior quality in the generated audio samples, as they imply reduced distortion and greater similarity to the original samples. On the other hand, a higher MOS (↑) signifies a more favorable human perception of the generated audio quality. It is noteworthy that the MOS metric also finds utility in evaluating the quality of synthesized videos, showcasing its versatility.

As presented in Tables 1–4, these are the experimental results in terms of these evaluation metrics using different datasets, which offer insights into the performance of our proposed approach as compared to the SOTA models. These results collectively provide a comprehensive understanding of the efficacy and advancements offered by our CCLCap-AE-AVSS framework.

Table 1 displays a thorough analysis of various VC models' performance on the VoxCeleb2 dataset. The evaluation of generated audio quality revolves around three main metrics: MCD, MSD, and MOS. Our model's performance is compared against several prominent VC models, namely, StarGAN-VC [44], StarGAN-VC2 [45], Blow [46], MelGAN-VC [47], FLSGAN-VC [48], and Exemplar AE [8]. Results demonstrate that our CCLCap-AE model excels in both MCD and MSD, achieving scores of 5.30 and 1.54, respectively. These scores notably outperform those of the alternative models, highlighting the superiority of our approach. Moreover, MOS values further support the effectiveness of our proposed approach over the SOTA models, as illustrated in Table 1.

Table 2 presents a comprehensive analysis of various VC models using the LRS3-TED dataset, focusing deliberately on the same SOTA models listed in Table 1. Our experiment demonstrates that our proposed CCLCap-AE model achieves outstanding performance, with MCD and MSD scores of 5.66 and 1.69, respectively. These scores are notably the lowest, indicating the effectiveness of our CCLCap-AE model in replicating the essential vocal characteristics of the target speaker. Additionally, our CCLCap-AE model receives the highest

**2** The generated samples are shared at: https://drive.google.com/drive/folders/1gXOtzn4pektByovO5ZA4qJb5H_qsxSI-?usp=sharing.

**Table 3:** Performance evaluation of VC models using VCTK dataset

| Model | MCD (↓) | MSD (↓) | MOS (↑) |
|---|---|---|---|
| StarGAN-VC [44] | 6.45 | 2.02 | 2.13 |
| StarGAN-VC2 [45] | 6.02 | 1.92 | 2.78 |
| Blow [46] | 5.91 | 1.79 | 2.92 |
| MelGAN-VC [47] | 6.11 | 1.97 | 2.64 |
| FLSGAN-VC [48] | 5.83 | 1.92 | 3.24 |
| Exemplar AE [8] | 5.65 | 1.71 | 3.10 |
| **CCLCap-AE (Ours)** | **5.49** | **1.66** | **3.68** |

**Table 4:** MOS (↑) of AVS models using VoxCeleb2 and LRS3-TED datasets

| Model | VoxCeleb2 | LRS3-TED |
|---|---|---|
| Speech2Vid [49] | 1.98 | 2.11 |
| LipGAN [50] | 2.74 | 2.68 |
| Exemplar AE [8] | 2.83 | 2.91 |
| **CCLCap-AE (Ours)** | **4.02** | **3.82** |

MOS value among all considered SOTA models, underscoring the superiority of our approach compared to other existing VC models. This observation is particularly significant, highlighting the excellence inherent in our method.

In addition, we have conducted an extensive evaluation of the performance of the VC component of our newly introduced CCLCap-AE model using VCTK dataset too. Upon careful examination of the results presented in Table 3, it becomes readily apparent that our model's MCD and MSD values stand out as the most optimal when compared to all other models under consideration. Furthermore, our model exhibits notably greater MOS values in comparison with its counterparts. The elevated MOS value serves as a clear indicator of the superior quality and fidelity of the generated samples achieved by our proposed framework.

Table 4 displays the MOS achieved by the AVS component of our proposed CCLCap-AE model and the SOTA models. In our assessment, we have taken into account three prominent SOTA models: Speech2Vid [49], LipGAN [50], and Exemplar AE [8]. The MOS values have been meticulously computed employing both the VoxCeleb2 and LRS3-TED datasets. The calculated MOS values for our AVS model stand at 4.02 and 3.82 for the VoxCeleb2 and LRS3-TED datasets, respectively. These results notably surpass the performance of all the aforementioned SOTA models. As demonstrated in Tables 1–4, the comparisons of both objective and subjective evaluations underscore the clear superiority of the proposed CCLCap-AE-AVSS over all the SOTA models in terms of both the audio and video qualities. This outcome highlights the remarkable performance of CCLCap-AE-AVSS in faithfully reproducing content and maintaining a high level of similarity to the input data, setting it apart as a leading model in the field.

Finally, the examination shifts toward visual inspection aided by mel-spectrograms. These mel-spectrograms depict the characteristics of the initial speech, which is taken into account for both the Exemplar AE and CCLap-AE models. Additionally, the Mel-spectrograms of the transformed speech segments generated by the Exemplar AE and CCLCap-AE models are also illustrated in Figure 5.

Upon scrutinizing the mel-spectrograms, it becomes evident that a discernible alteration in frequency components exists within the transformed speech instances of the Exemplar AE model. Conversely, the mel-spectrogram of the transformed speech resulting from CCLCap-AE closely mirrors that of the original speech, as shown in the marked area of Figure 5. This observation suggests a heightened resemblance between the converted speech and the original speech when employing the CCLCap-AE model.

Additionally, we provide visual depictions of both the source video and the videos created by Exemplar-AE and CCLCap-AE. This visual representation aids in conducting a thorough comparison concerning the

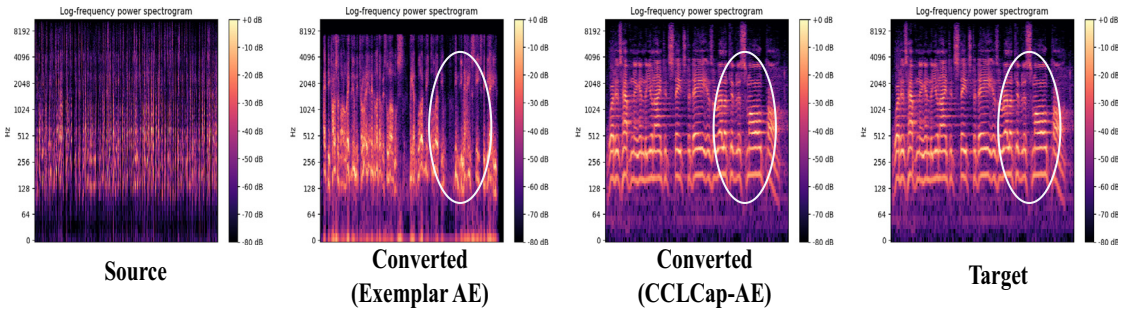


**Figure 5:** Visual comparison of the mel-spectrograms for both Exemplar-AE and CCLCap-AE with target mel-spectrograms.

authenticity and caliber of the produced video content, as illustrated in Figure 6. Through meticulous observation, it is apparent that the motion of the lips depicted in videos crafted by CCLCap-AE stands out distinctly with remarkable clarity. Conversely, upon scrutinizing videos generated by Exemplar-AE, a noticeable deficiency in effectively representing lip movements becomes evident. A direct juxtaposition with the source video underscores CCLCap-AE's superiority in video synthesis, manifesting in superior quality and naturalness. This model adeptly reproduces the original content with fidelity, yielding outcomes that are notably more realistic and authentic in comparison with those produced by Exemplar-AE.

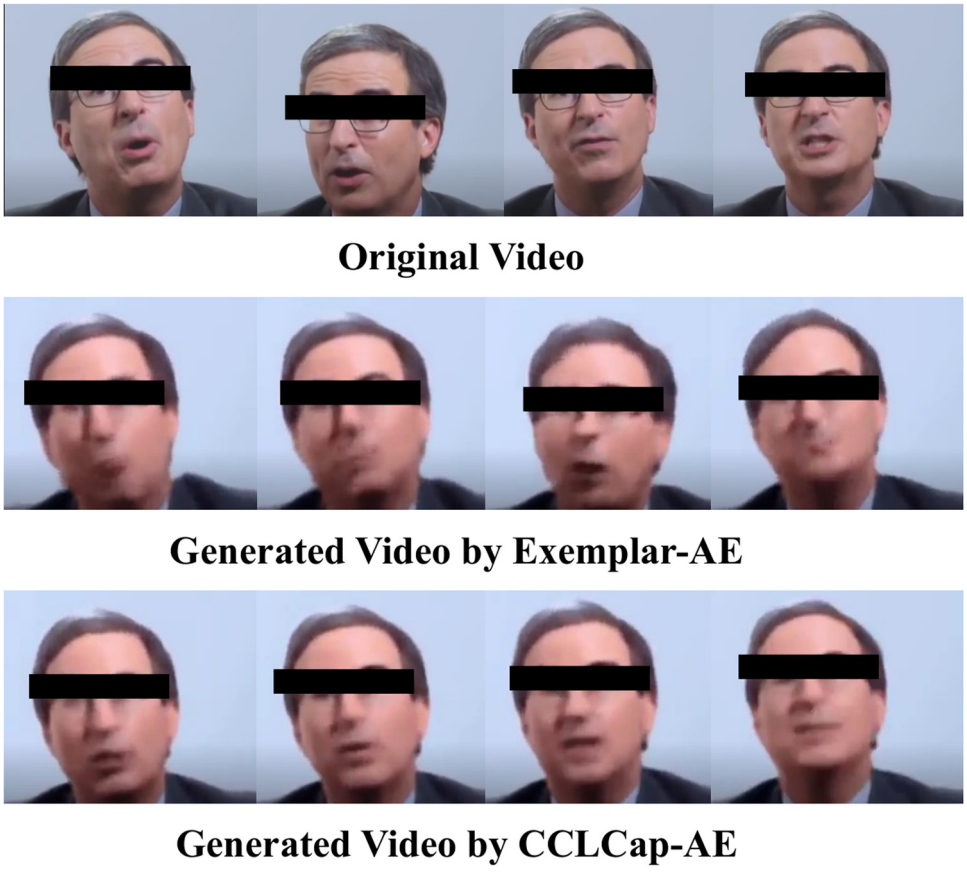


**Figure 6:** Visual comparison of the generated audio–visual stream for both Exemplar-AE and CCLCap-AE generated videos with the original video.

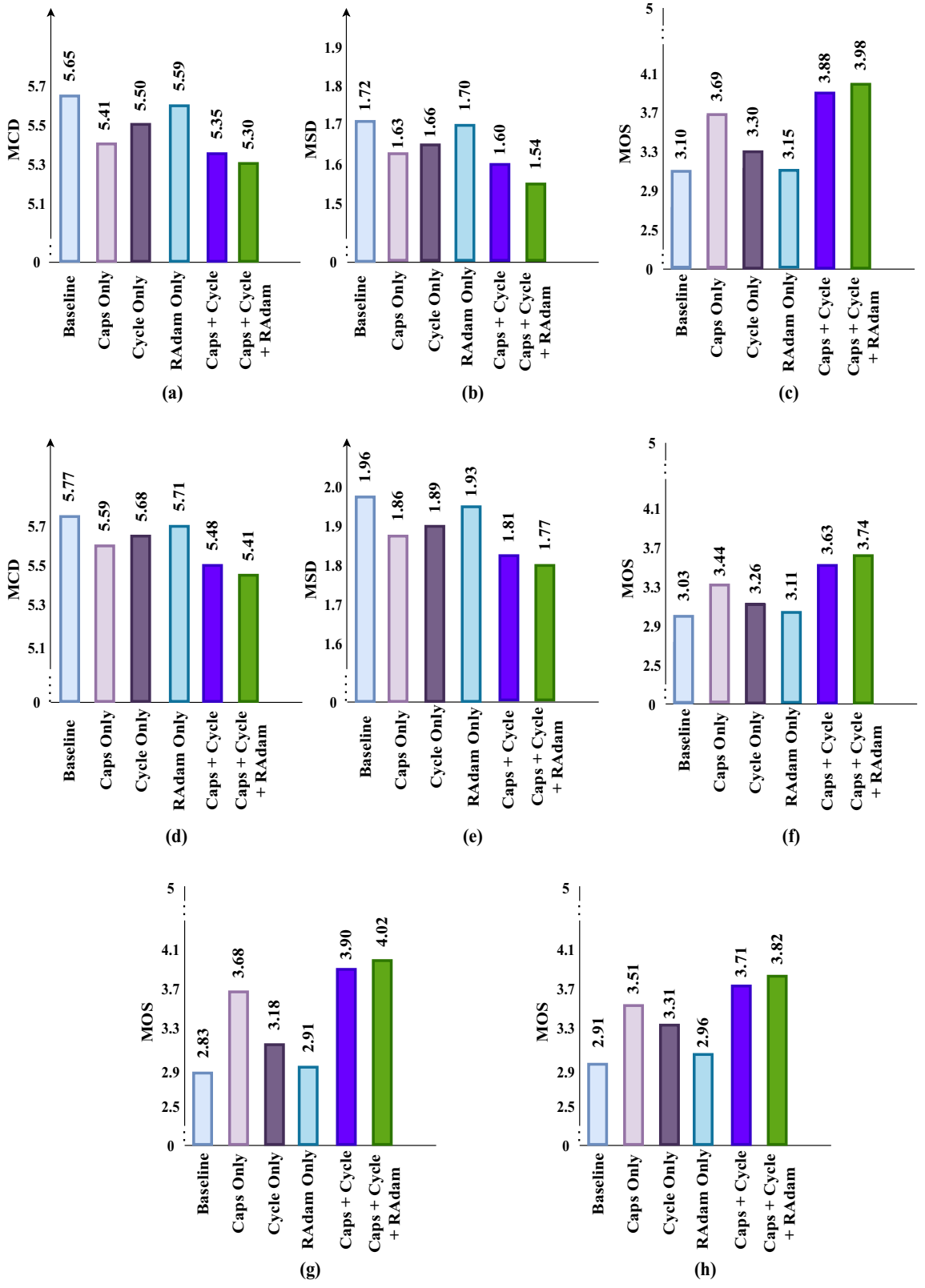


**Figure 7:** Ablation study of CCLCap-AE-AVSS: MCD, MSD, and MOS for audio and MOS for video using VoxCeleb2 and LRS3-TED datasets. (a) MCD for Audio (VoxCeleb2), (b) MSD for Audio (VoxCeleb2), (c) MOS for Audio (VoxCeleb2), (d) MCD for Audio (LRS3-TED), (e) MSD for Audio (LRS3-TED), (f) MCD for Audio (LRS3-TED), (g) MOS for Audio (VoxCeleb2), and (h) MOS for Audio (LRS3-TED).

### 6.2 Ablation study

We conducted an ablation study of the proposed CCLCap-AE-AVSS, delving into its inner workings by employing a range of evaluation metrics. These metrics included MCD, MSD, and MOS applied to the audio samples. Furthermore, we employed the MOS for the video samples. This comprehensive analysis involved the examination of six distinct scenarios, each with its own unique configuration. The scenarios we considered encompassed a variety of approaches: first, the baseline model [8]; second, the application of Caps-Net alone; third, the utilization of cycle consistency loss in isolation; fourth, the exclusive implementation of the RAdam optimizer; fifth, a combination of Caps-Net and cycle consistency loss; and finally, the proposed CCLCap-AE-AVSS, an integration of Caps-Net, cycle consistency loss, and RAdam. The outcomes of these scenarios are meticulously measured and visually represented in Figure 7, where the corresponding value of each metric has been prominently displayed at the top of its respective bar.

For the VoxCeleb2 dataset, we have obtained the MCD and MSD values for the generated speech of the CCLCap-AE model, which are determined to be 5.30 and 1.54, respectively. Similarly, when analyzing the LRS3-TED dataset, we have computed the MCD and MSD values and found them to be 5.41 and 1.77, respectively. These specific numerical results can be seen in panels (a), (b), (d), and (e) of Figure 7. Focusing on the MOS value, we observe that panels (c) and (f) of Figure 7 illustrate that the CCLCap-AE model achieves an MOS score of 3.98 for the VoxCeleb2 dataset and a score of 3.74 for the LRS3-TED dataset. Furthermore, to offer a more detailed breakdown, the respective values attributed to each component of the model are visually represented in Figure 7. For a comprehensive evaluation of the generated videos, we execute the MOS assessment on both the VoxCeleb2 and LRS3-TED datasets. The outcomes of these evaluations are depicted in panels (g) and (h) of Figure 7, where the CCLCap-AE attains an MOS value of 4.02 and 3.82 for the quality of generated audio and video, respectively.

The data showcased in our ablation study undeniably validate the superiority of our proposed CCLCap-AE-AVSS and each distinct model variant compared to the baseline model [8]. This marked enhancement can be attributed to the efficacy of the Caps-Net architecture in capturing salient features with precision. Additionally, the successful preservation of the original speech content's authenticity, achieved through the implementation of the cycle consistency loss, contributes significantly to the observed performance boost.

## 7 Limitations and future scope of works

While the CCLCap-AE-AVSS model shows promise in various areas, it is important to note some limitations. Our model's performance relies heavily on the availability and diversity of the training data. If the data lack diversity or do not adequately represent certain audio–visual speech patterns, the model's ability to generalize may suffer. Additionally, using Caps-Net architecture, especially with AEs, adds extra computational burden, resulting in longer training and inference times compared to traditional architectures. Furthermore, our model's robustness to noisy input, like imperfect audio recordings or challenging visual conditions, needs more attention. It may perform less optimally when faced with significant noise or input distortions.

To overcome these limitations in future research, we can explore techniques to enhance data augmentation strategies, optimize computational efficiency, and improve the model's ability to handle noisy input. Additionally, future research could investigate emotional VC and singing VC for AVSS. Furthermore, incorporating facial expressions corresponding to emotions (e.g., happiness, anger, sadness) into real-time video synthesis frameworks holds potential for advancing research in AVSS.

## 8 Conclusion

In summary, this research presents the CCLCap-AE-AVSS framework, a groundbreaking method for synthesizing audio–visual content linked to a specific target speaker through converted speech from a source speaker. The proposed approach integrates Caps-Net to effectively capture relevant features, and the cycle

consistency loss mechanism preserves the integrity of the source speaker's content. The rigorous evaluation demonstrates the CCLCap-AE-AVSS model's superior outputs in both auditory and visual domains compared to the SOTA models. Further exploration of advanced DL models, including GANs, holds promise for refining and advancing the AVSS process to enhance the performance and applicability of the proposed method in future research.

**Funding information**: The authors did not receive any funding from their institutes for this research.

**Author contributions:** Subhayu Ghosh and Nanda Dulal Jana contributed to the study conception and design. Coding implementation and analysis were performed by Subhayu Ghosh. The first draft of the manuscript was written by Subhayu Ghosh and Nanda Dulal Jana. Tapas Si, Saurav Mallik, and Mohd. Asif Shah participated in the revision of the paper and provided many pertinent suggestions. Nanda Dulal Jana guided throughout the research and supervised the every aspect of this work.

**Conflict of interest**: There is no conflict of interest to disclose.

**Data availability statement:** The source code of the proposed CCLCap-AE-AVSS is publicly available at: https://github.com/Subhayu-ghosh/CCLCap-AE-AVSS. The VoxCeleb2 and LRS3-TED datasets are used in this work, which can be found at: https://rb.gy/o7xs74 and https://rb.gy/5shq0j.

## References

[1] Desai S, Raghavendra EV, Yegnanarayana B, Black AW, Prahallad K. Voice conversion using artificial neural networks. In: 2009 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE; 2009. p. 3893–6.
[2] Mohammadi SH, Kain A. An overview of voice conversion systems. Speech Commun. 2017;88:65–82.
[3] Sisman B, Yamagishi J, King S, Li H. An overview of voice conversion and its challenges: From statistical modeling to deep learning. IEEE/ACM Trans Audio Speech Language Process. 2020;29:132–57.
[4] Cotescu M, Drugman T, Huybrechts G, Lorenzo-Trueba J, Moinet A. Voice conversion for whispered speech synthesis. IEEE Signal Process Lett. 2019;27:186–90.
[5] Barbulescu A, Hueber T, Bailly G, Ronfard R. Audio–visual speaker conversion using prosody features. In: AVSP 2013-12AVSP 2013-12th International Conference on Auditory-Visual Speech Processing; 2013. p. 11–6.
[6] Zhu H, Luo MD, Wang R, Zheng AH, He R. Deep audio–visual learning: A survey. Int J Automat Comput. 2021;18:351–76.
[7] Shi Z. A survey on audio synthesis and audio–visual multimodal processing. 2021. arXiv: http://arXiv.org/abs/arXiv:210800443.
[8] Deng K, Bansal A, Ramanan D. Unsupervised audiovisual synthesis via exemplar autoencoders. In: International Conference on Learning Representations; 2021.
[9] Bank D, Koenigstein N, Giryes R. Autoencoders. 2020. arXiv: http://arXiv.org/abs/arXiv:200305991.
[10] Zhai J, Zhang S, Chen J, He Q. Autoencoder and its various variants. In: 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE; 2018. p. 415–9.
[11] Jaiswal A, AbdAlmageed W, Wu Y, Natarajan P. Capsulegan: Generative adversarial capsule network. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops; 2018.
[12] Pande S, Chetty MSR. Analysis of capsule network (Capsnet) architectures and applications. J Adv Res Dynam Control Syst. 2018;10(10):2765–71.
[13] Wang R, Yang Z, You W, Zhou L, Chu B. Fake face images detection and identification of celebrities based on semantic segmentation. IEEE Signal Process Lett. 2022;29:2018–22.
[14] Kaneko T, Kameoka H. CycleGAN-VC: Non-parallel voice conversion using cycle-consistent adversarial networks. 2018 26th European Signal Processing Conference (EUSIPCO); 2018. p. 2100–4. doi: 10.23919/EUSIPCO.2018.8553236.
[15] Fang F, Yamagishi J, Echizen I, Lorenzo-Trueba J. High-quality nonparallel voice conversion based on cycle-consistent adversarial network. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE; 2018. p. 5279–83. doi: 10.1109/ICASSP.2018.8462342.
[16] Lorenzo-Trueba J, Fang F, Wang X, Echizen I, Yamagishi J, Kinnunen T. Can we steal your vocal identity from the Internet?: Initial investigation of cloning Obamaas voice using GAN, WaveNet and low-quality found data. 2018. The Speaker and Language Recognition Workshop, Speaker Odyssey 2018, doi: 10.21437/Odyssey.2018-34.
[17] Mazumder A, Ghosh S, Roy S, Dhar S, Jana ND. Rectified Adam Optimizer-Based CNN Model for Speaker Identification. In: Advances in Intelligent Computing and Communication: Proceedings of ICAC 2021. Springer; 2022. p. 155–62.

[18] Chung JS, Nagrani A, Zisserman A. Voxceleb2: Deep speaker recognition. 2018. arXiv: http://arXiv.org/abs/arXiv:180605622.
[19] Afouras T, Chung JS, Zisserman A. LRS3-TED: a large-scale dataset for visual speech recognition. 2018. arXiv: http://arXiv.org/abs/arXiv:180900496.
[20] Veaux C, Yamagishi J, MacDonald K. Superseded-CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit. 2019. The Centre for Speech Technology Research (CSTR), University of Edinburgh. doi: 10.7488/ds/2645.
[21] Makhzani A, Shlens J, Jaitly N, Goodfellow I, Frey B. Adversarial autoencoders. 2015. arXiv: http://arXiv.org/abs/arXiv:151105644.
[22] Zhang G, Liu Y, Jin X. A survey of autoencoder-based recommender systems. Front Comput Sci. 2020;14:430–50.
[23] Creswell A, White T, Dumoulin V, Arulkumaran K, Sengupta B, Bharath AA. Generative adversarial networks: an overview. IEEE Signal Process Magazine. 2018;35(1):53–65.
[24] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. Commun ACM. 2020;63(11):139–44.
[25] Sabour S, Frosst N, Hinton GE. Dynamic routing between capsules. Advances in Neural Information Processing Systems. USA: MIT Press, vol. 30; 2017.
[26] Vijayakumar T. Comparative study of capsule neural network in various applications. J Artif Intelligence. 2019;1(01):19–27.
[27] Patrick MK, Adekoya AF, Mighty AA, Edward BY. Capsule networks-a survey. J King Saud Univ Cmput Inform Sci. 2022;34(1):1295–310.
[28] Akhter MT, Banerjee P, Dhar S, Ghosh S, Jana ND. Region normalized capsule network based generative adversarial network for non-parallel voice conversion. In: International Conference on Speech and Computer. Springer Publication; 2023. p. 233–44.
[29] Toda T, Saruwatari H, Shikano K. Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum. In: 2001 IEEE international Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat No. 01CH37221). vol. 2. IEEE; 2001. p. 841–4.
[30] Kim EK, Lee S, Oh YH. Hidden Markov model based voice conversion using dynamic characteristics of speaker. In: European Conference On Speech Communication And Technology. Eurospeech; 1997. p. 2519–22.
[31] Toda T, Chen LH, Saito D. The voice conversion challenge 2016. In: Interspeech. USA: International Speech Communication Association (ISCA); 2016. p. 1632–6.
[32] Kaneko T, Kameoka H. Cyclegan-VC: non-parallel voice conversion using cycle-consistent adversarial networks. In: 2018 26th European Signal Processing Conference (EUSIPCO). IEEE; 2018. p. 2100–4.
[33] Sisman B, Zhang M, Dong M, Li H. On the study of generative adversarial networks for cross-lingual voice conversion. In: 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE; 2019. p. 144–51.
[34] Cassell J, Pelachaud C, Badler N, Steedman M, Achorn B, Becket T, et al. Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. In: Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques; 1994. p. 413–20.
[35] Sawada K, Takehara M, Tamura S, Hayamizu S. Audio–visual voice conversion using noise-robust features. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE; 2014. p. 7899–903.
[36] Moubayed SA, Smet MD, Van Hamme H. Lip synchronization: from phone lattice to PCA eigen-projections using neural networks. In: Ninth Annual Conference of the International Speech Communication Association. Citeseer; 2008.
[37] Ibrokhimov B, Hur C, Kim H, Kang S. A-DBNF: adaptive deep belief network framework for regression and classification tasks. Appl Intelligence. 2021;51(7):4199–213.
[38] Hardoon DR, Szedmak S, Shawe-Taylor J. Canonical correlation analysis: An overview with application to learning methods. Neural Comput. 2004;16(12):2639–64.
[39] Tamura S, Horio K, Endo H, Hayamizu S, Toda T. Audio–visual voice conversion using deep canonical correlation analysis for deep Bottleneck features. In: INTERSPEECH. India: International Speech Communication Association (ISCA); 2018. p. 2469–73.
[40] Redfern J. Video to audio conversion for the visually impaired. School of Computer Science & Informatics, Cardiff University, (May 2015); 2015.
[41] Durak L, Arikan O. Short-time Fourier transform: two fundamental properties and an optimal implementation. IEEE Trans Signal Process. 2003;51(5):1231–42.
[42] Hwang Y, Cho H, Yang H, Won DO, Oh I, Lee SW. Mel-spectrogram augmentation for sequence to sequence voice conversion. 2020. arXiv: http://arXiv.org/abs/arXiv:200101401.
[43] Wang SL, Lau WH, Liew AWC, Leung SH. Robust lip region segmentation for lip images with complex background. Pattern Recognition. 2007;40(12):3481–91.
[44] Kameoka H, Kaneko T, Tanaka K, Hojo N. Stargan-VC: non-parallel many-to-many voice conversion using star generative adversarial networks. In: 2018 IEEE Spoken Language Technology Workshop (SLT). IEEE; 2018. p. 266–73.
[45] Kaneko T, Kameoka H, Tanaka K, Hojo N. Stargan-vc2: Rethinking conditional methods for stargan-based voice conversion. 2019. arXiv: http://arXiv.org/abs/arXiv:190712279.
[46] Serrà J, Pascual S, Segura Perales C. Blow: a single-scale hyperconditioned flow for non-parallel raw-audio voice conversion. Advances in Neural Information Processing Systems. USA: MIT Press; vol. 32. 2019.
[47] Pasini M. MelGAN-VC: voice conversion and audio style transfer on arbitrarily long samples using spectrograms. 2019. arXiv: http://arXiv.org/abs/arXiv:191003713.

[48] Dhar S, Banerjee P, Jana ND, Das S. Voice conversion using feature specific loss function based self-attentive generative adversarial network. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE; 2023. p. 1–5.
[49] Chung JS, Jamaludin A, Zisserman A. You said that? 2017. arXiv: http://arXiv.org/abs/arXiv:170502966.
[50] Prajwal KR, Mukhopadhyay R, Philip J, Jha A, Namboodiri V, Jawahar C. Towards automatic face-to-face translation. In: Proceedings of the 27th ACM International Conference on Multimedia; 2019. p. 1428–36.
[51] Tobing PL, Wu YC, Hayashi T, Kobayashi K, Toda T. Non-parallel voice conversion with cyclic variational autoencoder. 2019. arXiv: http://arXiv.org/abs/arXiv:190710185.
[52] Akhter MT, Banerjee P, Dhar S, Jana ND. An analysis of performance evaluation metrics for voice conversion models. In: 2022 IEEE 19th India Council International Conference (INDICON). IEEE; 2022. p. 1–6.
[53] Tang Y, Cooke M. Subjective and objective evaluation of speech intelligibility enhancement under constant energy and duration constraints. In: Twelfth Annual Conference of the International Speech Communication Association; 2011.
[54] Kubichek R. Mel-cepstral distance measure for objective speech quality assessment. In: Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing. vol. 1. IEEE; 1993. p. 125–8.
[55] Takamichi S, Toda T, Black AW, Neubig G, Sakti S, Nakamura S. Postfilters to modify the modulation spectrum for statistical parametric speech synthesis. IEEE/ACM Trans Audio Speech Language Process. 2016;24(4):755–67.
[56] Streijl RC, Winkler S, Hands DS. Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives. Multimedia Systems. 2016;22(2):213–27.