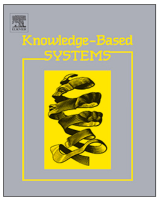
Knowledge-Based Systems 277 (2023) 110837

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys



End-to-end translation of human neural activity to speech with a dual–dual generative adversarial network

Yina Guo [a,*,1], Ting Liu [a,1], Xiaofei Zhang [b,1], Anhong Wang [a], Wenwu Wang [c]

[a] *School of Electronic Information Engineering, Taiyuan University of Science and Technology, Taiyuan, 030024, Shanxi, China*
[b] *Shanxi Jiashida Robot Technology Co. LTD, Taiyuan, 030006, Shanxi, China*
[c] *Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, GU2 7XH, Surrey, UK*

ARTICLE INFO

Article history:
Received 7 October 2022
Received in revised form 20 July 2023
Accepted 21 July 2023
Available online 25 July 2023

Dataset link: https://github.com/qwe1218088/dual-dualgan.git

Keywords:
Translation of human neural activity to speech (T-CAS)
End-to-end model
Dual–dual generative adversarial network (dual-dualGAN)
Brain–computer interface (BCI)
Cross-domain mapping

ABSTRACT

In a recent study of auditory evoked potential (AEP) based brain–computer interface (BCI), it was shown that, with an encoder–decoder framework, it is possible to translate human neural activity to speech (T-CAS). Current encoder–decoder-based methods achieve T-CAS often with a two-step approach where the information is passed between the encoder and decoder with a shared vector of reduced dimension, which, however, may result in information loss. In this paper, we propose an end-to-end model to translate human neural activity to speech (ET-CAS) by introducing a dual–dual generative adversarial network (Dual-DualGAN) for cross-domain mapping between electroencephalogram (EEG) and speech signals. In this model, we bridge the EEG and speech signals by introducing transition signals which are obtained by cascading the corresponding EEG and speech signals in a certain proportion. We then learn the mappings between the speech/EEG signals and the transition signals. We also develop a new EEG dataset where the attention of the participants is detected before the EEG signals are recorded to ensure that the participants have good attention in listening to speech utterances. The proposed method can translate word-length and sentence-length sequences of neural activity to speech. Experimental results show that the proposed method significantly outperforms state-of-the-art methods on both words and sentences of auditory stimulus.

1. Introduction

The World Health Organization (WHO) estimated in 2021 that neurological disorders could affect as many as 25% patients worldwide, and result in symptoms including confusion, altered levels of consciousness, and loss of communication. The visual evoked potential (VEP) based brain–computer interface (BCI) may enhance the quality of life of a patient, e.g. by using eyes to control a cursor for selecting letters one-by-one to spell out words [1–3]. However, the spelling rates of users are far below the average rate of 150 words per minute in natural speech [4,5], since spelling is a sequential concatenation of discrete letters [6–8]. Different from spelling, speech is a highly efficient form of communication produced from a fluid stream of overlapping and multi-articulator vocal tract movements [9–11]. The auditory evoked potential (AEP) based BCI is a promising alternative to overcome the limitations of current spelling-based methods in achieving natural communication rates [12–17].

The AEP based BCI using spelling-based methods, however, are yet to reach natural communication rates. To address this problem, studies have been conducted to exploit the conceptual similarity between the task of decoding speech from human neural activity and the task of machine translation according to the sensitivity of organs (such as ears, eyes). The AEP based BCI methods can be mainly classified into two categories: translation of human neural activity to text (T-CAT) [18–20] and translation of human neural activity to speech (T-CAS) [15,21–30]. The T-CAT method is mainly used by deaf-mutes. However, this method can be limited in several scenarios. For example, when two words share the same pronunciation, the translation to the desired word can be ambiguous. In addition, the spelling rates achieved by T-CAT can only be close to able-bodied typing rates. In contrast, the T-CAS method can be used by more users, and an increasing number of neural network based on this approach have been proposed [31,32]. It is a more intuitive approach for communication, as in natural speech [16,17,33–35]. As a result, the T-CAS approach has received increasing interest recently, and is the focus of this paper.

In existing AEP based BCIs, the T-CAS is often achieved with a two-step method, in which the first step is to encode neural activities to texts or acoustic features, typically for dimension

* Corresponding author.
E-mail addresses: zulibest@tyust.edu.cn (Y. Guo), w.wang@surrey.ac.uk (W. Wang).

[1] Contributed equally to this work.

https://doi.org/10.1016/j.knosys.2023.110837
0950-7051/© 2023 Elsevier B.V. All rights reserved.

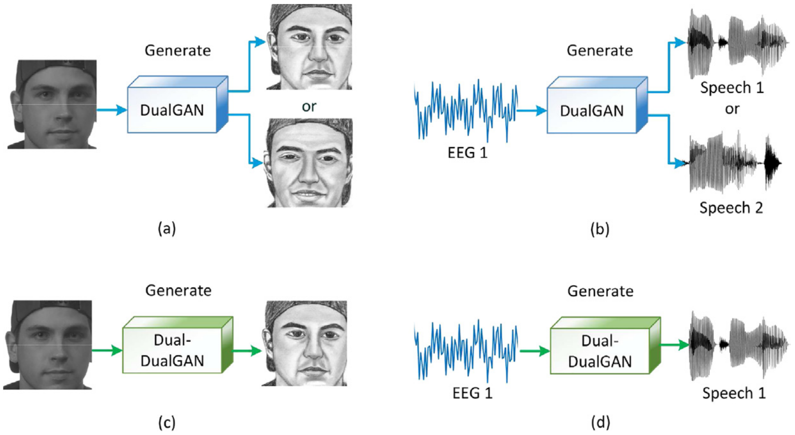


Fig. 1. Translation results of a DualGAN and our end-to-end model (Dual-DualGAN). (a) Image-to-image translation by a DualGAN, (b) EEG-to-speech translation by a DualGAN, (c) Image-to-image translation by a Dual-DualGAN, and (d) EEG-to-speech translation by a Dual-DualGAN.

reduction, followed by a second step on decoding texts or acoustic features to synthesized speech by a vocoder or a speech synthesis tool. In the two-step method, a dimension reduced vector is often used to bridge the decoder and encoder, but dimension reduction may lead to the information loss and reconstruction error [36,37], which may result in inaccurate signal reconstruction by using different decoders, such as [16,17,33,35]. To our knowledge, there is no existing study for end-to-end decoding of human neural activity to speech, by using non-invasive electroencephalogram (EEG) neural recordings, without involving dimension reduction as in the two-step method.

The aim of this paper is to develop an end-to-end method for translating human neural activity to speech. To this end, we leverage the dual generative adversarial network (DualGAN) [38], which can generate paired images with same features, and does not involve dimension reduction in the pipeline. This offers advantages over the cycle-consistent adversarial networks (CycleGAN) [39] or the denoising diffusion probabilistic models (DDPM) [40]. More specifically, CycleGAN was designed for unpaired image-to-image translation but not the generation of paired images with same features, while DDPM is a progressive lossy decompression scheme trained using variational inference to produce samples matching the data distribution using multiple steps with much slower sampling speed than GAN.

The DualGAN, however, may be limited by the following challenges. For example, it is an unsupervised dual learning framework originally designed for cross-domain image-to-image translation, but it cannot achieve a one-to-one translation for different kind of signal pairs, such as EEG and speech signals, due to the lack of corresponding features between these modalities. As shown in Fig. 1(a), a male photo may be translated to the corresponding male sketch or other similar male sketches by DualGAN. This is because the image pairs have commonalities, for example, hat or hair. In Fig. 1(b), an EEG signal may be translated to different speech signals randomly by a DualGAN, as the EEG and speech signals have different dimensions, without corresponding features (e.g. waveform and amplitude) over time. To address these challenges, we propose an end-to-end model for the translation of human neural activity to speech (ET-CAS). Our contributions are three-fold:

1. **Model.** An end-to-end model is proposed to translate human neural activity to speech directly.
2. **Datasets.** A new EEG dataset is created for this study, where a device (see Fig. 7) is designed to detect the attention of participants in order to improve the quality of the EEG data in data collection.
3. **Network.** A dual–dual generative adversarial network (Dual-DualGAN) is proposed to address the ET-CAS problem, where two DualGANs are built and trained simultaneously by incorporating a transition domain to bridge the two DualGANs. The transition signals used in the transition domain are obtained by cascading the corresponding EEG and speech signals in a certain proportion, where shared labels are constructed for EEG and speech signals without aligning their corresponding features.

The remainder of the paper is organized as follows. Section 2 describes the related work. Section 3 introduces the background for GAN and DualGAN. Section 4 formulates an end-to-end model for the ET-CAS problem. Section 5 presents our proposed network for the problem of ET-CAS. Section 6 describes data collection and pre-processing. Section 7 discusses the experimental set up. Section 8 shows numerical results. Section 9 concludes the paper and draws potential future research directions.

2. Related work

In the field of the AEP based BCIs, there is an increasing interest in the problem of decoding speech from human neural activity [16,17,33–35]. According to the sensitivity of organs (e.g. ears, eyes), the AEP based BCI systems can be mainly classified into two categories, namely, T-CAT and T-CAS. The T-CAT systems are more suitable for deaf and mute people, and the spelling rates offered by these systems are close to typing rates. However, they are prone to errors when two words are translated with the same pronunciation. In contrast, the T-CAS method does not have such limitations, and is an intuitive approach for users to achieve high communication rates as in natural speech.

The AEP based BCIs for T-CAT. Herff et al. [41] showed for the first time that spoken speech could be decoded into the expressed words from intracranial electrocorticographic (ECoG) recordings, and proposed a Brain-To-Text system to transform brain activity while speaking into the corresponding textual representation. This system achieved word error rates as low as 25% and phone error rates below 50%. Makin et al. [19] trained a recurrent neural network to encode each sentence-long sequence of neural activity into an abstract representation, and then to decode this representation, word by word, into an English sentence at natural-speech rates with high accuracy. This method achieved an average word error rate across a held-out repeat set as low as 3%. Willett et al. [20] proposed a BCI which can spell 90 characters per minute at >99% accuracy with general-purpose auto-correction, and significantly close the gap between BCI-enabled typing and able-bodied

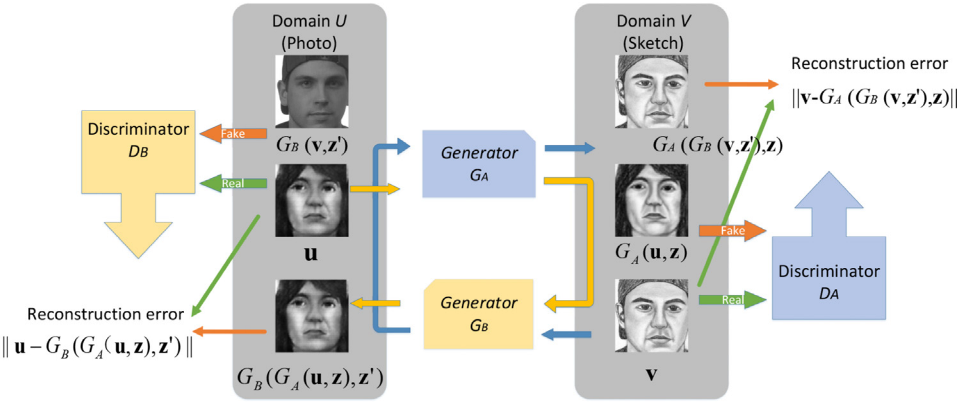


Fig. 2. Architecture and data flow chart of DualGAN for image-to-image translation.

typing rates. These methods have focused on the translation of human neural activity to text and achieved typing rates that are close to normal typing rates, which, however, remain lower than the average rate of 150 words per min in natural speech.

The AEP based BCIs for T-CAS. Brumberg et al. [42] developed a brain–computer interface for the control of an artificial speech synthesizer by an individual with near complete paralysis, where vowel formant frequencies are predicted based on neural activity recorded from an intra-neural micro-electrode implanted in the left hemisphere speech motor cortex. Bocquelet et al. [15] presented an articulator-based speech synthesizer, which converts movements of the main speech articulators (e.g. tongue, jaw, velum, and lips) into intelligible speech by using a deep neural network (DNN), and can be controlled in real-time, which is crucial for BCI applications. Akbari et al. [16] investigated the dependence of reconstruction accuracy on linear and nonlinear (deep neural network) regression methods and the acoustic representation. Anumanchipalli et al. [17] designed a neural decoder that explicitly leverages kinematic and sound representations encoded in human neural activity to synthesize audible speech. These methods demonstrated the possibility of translating human neural activity to speech with encoder–decoder frameworks. However, there are two major open challenges for T-CAS. First, the collection of intracranial ECoG recordings is intrusive and inconvenient. Second, the encoder–decoder based methods need multi-steps to achieve T-CAS. Krishna et al. [34] demonstrated synthesizing speech from the non-invasive electroencephalogram (EEG) neural recordings for the first time and proposed a recurrent neural network (RNN) regression model to predict mel-frequency cepstral coefficients (MFCC) from EEG features. This method shows that it is possible to decode the human neural activity via non-invasive EEG neural recordings, but does not consider speech reconstruction from the EEG features.

The focus of this paper is to address the problem of decoding speech from human neural activity directly with the non-invasive EEG signals, and to propose an end-to-end method Dual-DualGAN.

3. Background

A generative adversarial network (GAN) is a class of machine learning frameworks designed by Goodfellow et al. [43]. The first GAN architecture used fully connected neural networks for both the generator and discriminator. This was then extended by replacing the fully connected network with convolutional neural networks (CNN), which is well suited for image data [44,45]. Given a training set, a GAN learns to generate new data with the same statistics as the training set. For example, a GAN trained on photos can generate new photos that look at least superficially authentic to human observers, having many realistic characteristics. However, the original GAN algorithm is not directly applicable to the translation task, because the goal of the original GAN is to estimate the distribution of the training data (such as images) and generate new data from that distribution, while in translation tasks, this is often not the case.

Motivated by GAN and dual learning, an unsupervised learning framework DualGAN has been proposed by Yi et al. [38] for image translation, which was trained with two sets of unlabeled images from two domains (e.g. sketch and photo).

As illustrated in Fig. 2, given two sets of unlabeled and unpaired images sampled from domains U and V, respectively, the primary task of DualGAN is to learn a generator G_A: $U \rightarrow V$ that maps an image $\mathbf{u} \in U$ to an image $\mathbf{v} \in V$, while the dual task is to train an inverse generator G_B: $V \rightarrow U$. This is achieved by using two GANs (i.e. the primal GAN and the dual GAN). The primal GAN learns the generator G_A and a discriminator D_A that discriminates between fake outputs of G_A and real members of domain V. Analogously, the dual GAN learns the generator G_B and a discriminator D_B.

The image $\mathbf{u} \in U$ is translated to domain V using G_B. How well the translation $G_A(\mathbf{u}, \mathbf{z})$ fits in V is evaluated by D_A, where $\mathbf{z}$ is random noise. $G_A(\mathbf{u}, \mathbf{z})$ is then translated back to domain U using G_A, which outputs $G_B(G_A(\mathbf{u}, \mathbf{z}), \mathbf{z}')$ as the reconstructed version of $\mathbf{u}$, where $\mathbf{z}'$ is also random noise. Similarly, $\mathbf{v} \in V$ is translated to U as $G_B(\mathbf{v}, \mathbf{z}')$ and then reconstructed as $G_A(G_B(\mathbf{v}, \mathbf{z}'), \mathbf{z})$. The discriminator D_A is trained with $\mathbf{v}$ as positive examples and $G_A(\mathbf{u}, \mathbf{z})$ as negative examples, whereas D_B takes $\mathbf{u}$ as positive and $G_B(\mathbf{v}, \mathbf{z}')$ as negative. The generators G_A and G_B are optimized to emulate "fake" outputs to fool the corresponding discriminators D_A and D_B, as well as to minimize the two reconstruction losses $\|\mathbf{u} - G_B(G_A(\mathbf{u}, \mathbf{z}), \mathbf{z}')\|$ and $\|\mathbf{v} - G_A(G_B(\mathbf{v}, \mathbf{z}'), \mathbf{z})\|$.

The same loss function is used for both the generators G_A and G_B as they share the same task, which is defined as

$$\begin{aligned}\mathcal{L}(U, V)^G &= \lambda_U \|\mathbf{u} - G_B(G_A(\mathbf{u}, \mathbf{z}), \mathbf{z}')\| \\ &+ \lambda_V \|\mathbf{v} - G_A(G_B(\mathbf{v}, \mathbf{z}'), \mathbf{z})\| \\ &- D_A(G_A(\mathbf{u}, \mathbf{z})) - D_B(G_B(\mathbf{v}, \mathbf{z}')),\end{aligned} \tag{1}$$

where λ_U and λ_V are two constant parameters, which are typically set to the values within [100, 1000] [38].

The loss functions of D_A and D_B advocated by Wasserstein GAN (WGAN) [38] can be described by

$$\mathcal{L}_A^D = D_A(G_A(\mathbf{u}, \mathbf{z})) - D_A(\mathbf{v}), \tag{2}$$

$$\mathcal{L}_B^D = D_B(G_B(\mathbf{v}, \mathbf{z}')) - D_B(\mathbf{u}), \tag{3}$$

where $D_A(\cdot) = \frac{p_{data}(\cdot)}{p_{data}(\cdot)+p_{gA}(\cdot)}$, $D_B(\cdot) = \frac{p_{data}(\cdot)}{p_{data}(\cdot)+p_{gB}(\cdot)}$ $p_{data}(\cdot)$ is the distribution of the training data, $p_{gA}(\cdot)$ and $p_{gB}(\cdot)$ are the distributions

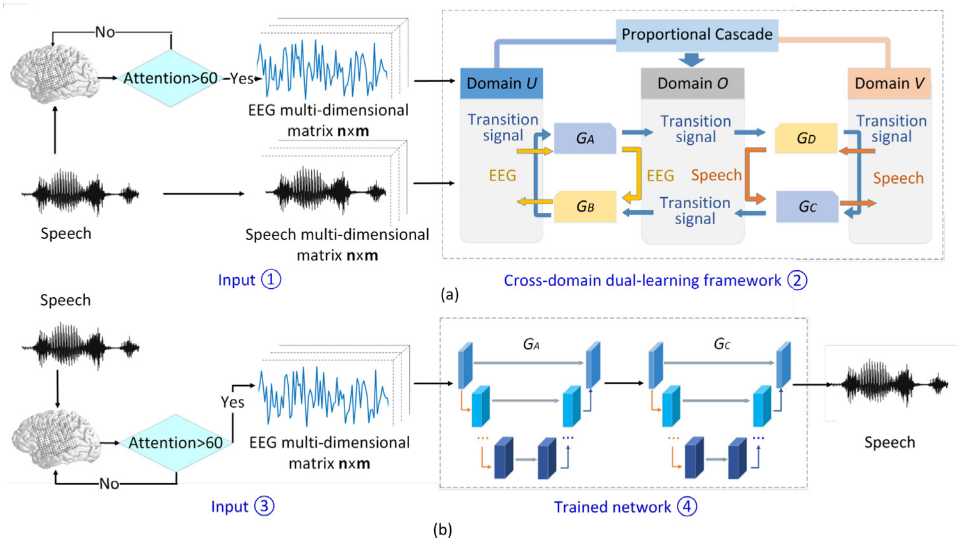


Fig. 3. Training and testing process of our end-to-end model (Dual-DualGAN) for EEG-to-speech translation.

of the fake outputs from $G_A(\cdot)$ and $G_B(\cdot)$, respectively. After several steps of training, if $p_g(\cdot) = p_{data}(\cdot)$, the discriminator is unable to differentiate between $p_g(\cdot)$ and $p_{data}(\cdot)$, and $D_A(\cdot) = D_B(\cdot) = \frac{1}{2}$.

By training a DualGAN, a signal can be translated to another similar signal with some correspondence in different patterns. For example, a male photo may be translated to the corresponding male sketch or other similar male sketches by a DualGAN in Fig. 1(a). However, the male photo and the translated male sketch are not necessarily image pairs. Analogously, an EEG signal may be translated to some different speech signals by a DualGAN in Fig. 1(b). The correct rates of the one-to-one translation of EEG-to-speech are even lower than those for image pairs without local corresponding features.

For the ET-CAS problem, we need to realize one-to-one translation of an EEG to a speech signal. To address this problem, in this paper, we build an end-to-end model for decoding speech from human neural activities. Based on the end-to-end model, we propose a Dual-DualGAN by group labeling the EEG and speech signals, and inserting a transition domain into the DualGAN to train two DualGANs simultaneously. The transition signals in the transition domain can be considered as the shared labels for the EEG and speech signals which are constructed by cascading the corresponding EEG and speech signals in a certain proportion.

4. End-to-end model

In this section, we present an end-to-end model for decoding speech from human neural activities, as illustrated in Fig. 3.

Training. Fig. 3(a) shows the training process of our end-to-end model. In this model, instead of using current encoder–decoder frameworks with two-step methods, we design a cross-domain dual-learning framework Dual-DualGAN to bridge EEG and speech signals without extracting their corresponding features, and realize one-to-one cross-domain EEG-to-speech translation. Two inputs are used in this framework, including the non-invasive EEG signal recorded as the participants listen to speech sound, and the corresponding speech signal. We use these two inputs to train the proposed Dual-DualGAN, and then obtain the trained Dual-DualGAN.

Testing. Fig. 3(b) demonstrates the testing process of the end-to-end model. The input of the model is the non-invasive EEG signal recorded as the participants listen to speech sound. The parameters derived from the trained Dual-DualGAN is used to synthesize the speech signal in terms of the corresponding EEG signal.

5. Dual-DualGAN

The fundamental question in ET-CAS is to decode speech from human neural activity (EEG signals) directly. A potential approach to this problem is to use a DualGAN which is an unsupervised dual learning framework that does not need dimensionality reduction in cross-modal translation. However, it cannot realize one-to-one translation for signal pairs without local corresponding features from the signal pairs. The EEG and speech signals are different types of signals without local corresponding features. Thus we cannot achieve one-to-one translation of an EEG to a speech signal by training a DualGAN. To address this problem, we present a Dual-DualGAN, as shown in Fig. 4.

General ideas. Dual-DualGAN involves a domain U, a domain V and a transition domain O. The domain O is introduced into a DualGAN to build a left mini-cycle, a right mini-cycle and a large cycle, which are trained simultaneously.

The sets of EEG signals $\mathbf{u}$ and speech signals $\mathbf{v}$ are sampled from domains U and V, respectively. As illustrated in Fig. 5(b), a set of transition signals $\mathbf{o}$ sampled from O are obtained by cascading the corresponding EEG and speech signals in a certain proportion.

The primary task of our Dual-DualGAN is to translate $\mathbf{u} \in U$ into $\mathbf{v} \in V$. The left mini-cycle aims to learn a mapping between the EEG signals $\mathbf{u} \in U$, while the right mini-cycle learns a mapping between the speech signals $\mathbf{v} \in V$. Different from the above running mode, the large cycle aims to learn a mapping between the transition signals $\mathbf{o} \in O$ and the EEG signals $\mathbf{u} \in U$ from O to U, and then to learn a mapping between the transition signals $\mathbf{o} \in O$ and the speech signals $\mathbf{v} \in V$ from O to V. By training the Dual-DualGAN, the transition signals $\mathbf{o} \in O$ can be considered as shared labels for EEG and speech signals without their corresponding features, with details discussed as follows.

5.1. Training

Left mini-cycle. In Fig. 4(a), for EEG signals, a real EEG signal $\mathbf{u}$ is mapped to domain O using a generator G_A: $U \rightarrow O$, which generates an EEG signal $G_A(\mathbf{u}, \mathbf{z})$. Then $G_A(\mathbf{u}, \mathbf{z})$ is translated back to domain U using an inverse generator G_B: $O \rightarrow U$, which outputs $G_B(G_A(\mathbf{u}, \mathbf{z}), \mathbf{z}')$ as the reconstructed version of $\mathbf{u}$.

Right mini-cycle. For speech signals, a real speech signal $\mathbf{v}$ is mapped to domain O using a generator G_C: $V \rightarrow O$, which generates a speech signal $G_C(\mathbf{v}, \mathbf{z}')$. $G_C(\mathbf{v}, \mathbf{z}')$ is then translated

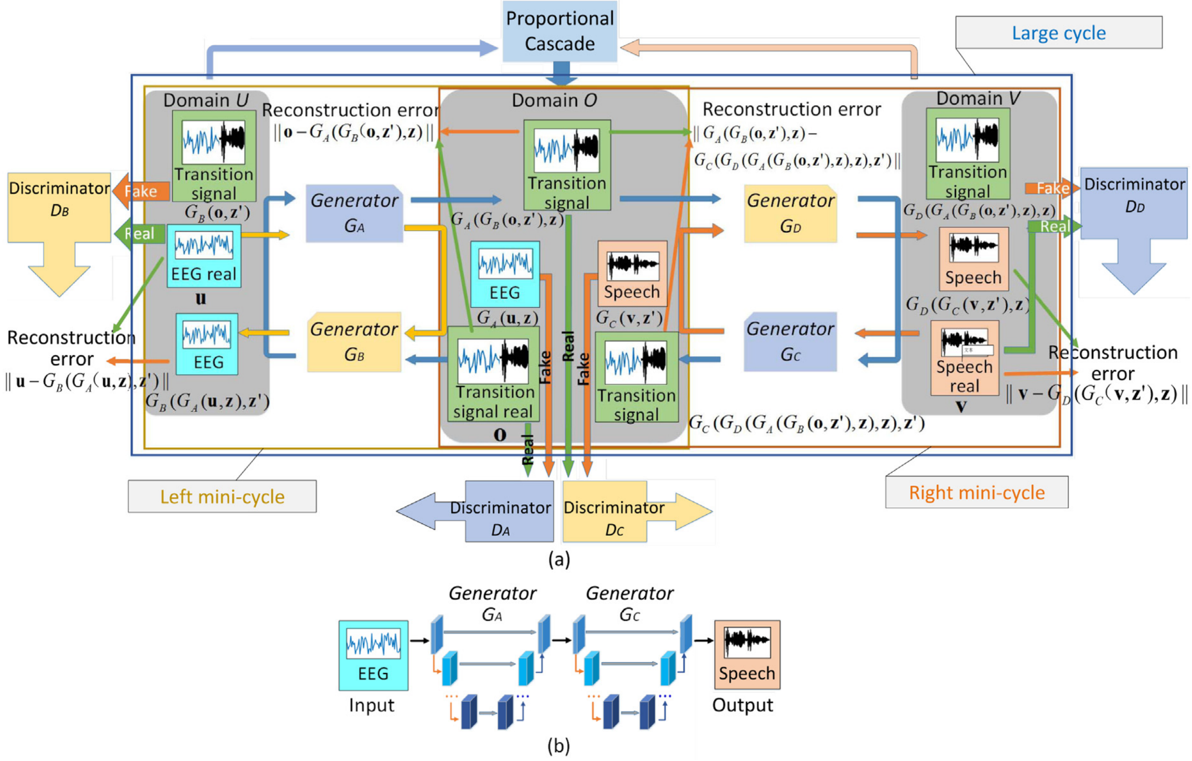

Proportional Cascade
Large cycle
Domain U
Domain O
Domain V
Reconstruction error
$\|\mathbf{o}-G_A(G_B(\mathbf{o},\mathbf{z}'),\mathbf{z})\|$
$\|G_A(G_B(\mathbf{o},\mathbf{z}'),\mathbf{z})-G_C(G_D(G_A(G_B(\mathbf{o},\mathbf{z}'),\mathbf{z}),\mathbf{z}),\mathbf{z}')\|$
Discriminator D_B
Fake
Real
Transition signal
$G_B(\mathbf{o},\mathbf{z}')$
EEG real
u
Generator G_A
Generator G_B
Generator G_D
Generator G_C
$G_A(G_B(\mathbf{o},\mathbf{z}'),\mathbf{z})$
EEG
Speech
$G_A(\mathbf{u},\mathbf{z})$
$G_C(\mathbf{v},\mathbf{z}')$
Transition signal real
o
Reconstruction error
$\|\mathbf{u}-G_B(G_A(\mathbf{u},\mathbf{z}),\mathbf{z}')\|$
$G_B(G_A(\mathbf{u},\mathbf{z}),\mathbf{z}')$
$G_D(G_A(G_B(\mathbf{o},\mathbf{z}'),\mathbf{z}),\mathbf{z})$
$G_D(G_C(\mathbf{v},\mathbf{z}'),\mathbf{z})$
Speech real
v
Discriminator D_D
$\|\mathbf{v}-G_D(G_C(\mathbf{v},\mathbf{z}'),\mathbf{z})\|$
$G_C(G_D(G_A(G_B(\mathbf{o},\mathbf{z}'),\mathbf{z}),\mathbf{z}),\mathbf{z}')$
Left mini-cycle
Right mini-cycle
Discriminator D_A
Discriminator D_C
(a)
Generator G_A
Generator G_C
EEG
Input
Speech
Output
(b)

Fig. 4. Architecture of our network (a) Dual-DualGAN for training, including two DualGANs trained simultaneously. DualGAN 1 learns the mapping between the EEG signals $\mathbf{u} \in U$ and the transition signals $\mathbf{o} \in O$, while DualGAN 2 learns the mapping between the speech signals $\mathbf{v} \in U$ and the generated transition signals $\mathbf{o} \in O$. Thus we can find the mapping between the EEG signals $\mathbf{u} \in U$ and the speech signals $\mathbf{v} \in V$ to address the ET-CAS problem. (b) The network used in testing, where the trained Dual-DualGAN is used to achieve EEG-to-speech translation.

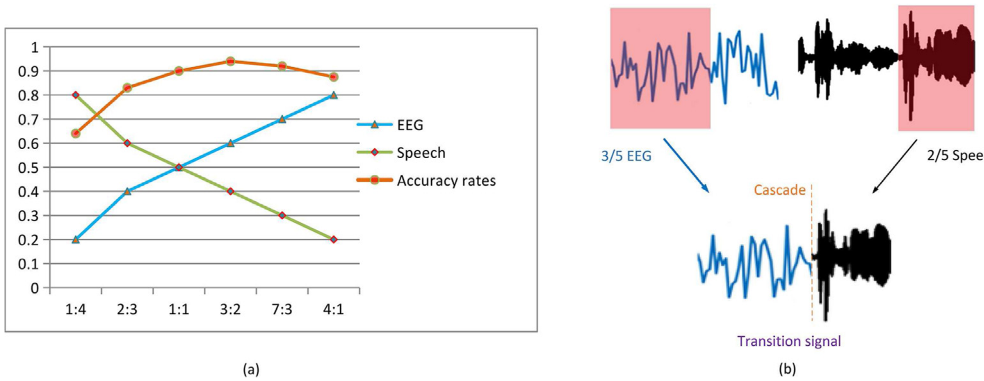


Fig. 5. Proportional cascade of the transition signals (a) Accuracy with respect to different cascading proportions (b) An example of the transition signal formed by cascading the EEG and speech signal in a ratio of 3:2.

back to domain V using an inverse generator G_D: $O \rightarrow V$, which outputs $G_D(G_C(\mathbf{v}, \mathbf{z}'), \mathbf{z})$ as the reconstructed version of $\mathbf{v}$.

Large cycle. For transition signals, it needs four steps to form a large cycle. Firstly, a real transition signal $\mathbf{o}$ is mapped to domain U using G_B, which generates a transition signal $G_B(\mathbf{o}, \mathbf{z}')$. Secondly, $G_B(\mathbf{o}, \mathbf{z}')$ is translated back to domain O using G_A, which outputs $G_A(G_B(\mathbf{o}, \mathbf{z}'), \mathbf{z})$. Thirdly, $G_A(G_B(\mathbf{o}, \mathbf{z}'), \mathbf{z})$ is translated to domain V using G_D, which generates $G_D(G_A(G_B(\mathbf{o}, \mathbf{z}'), \mathbf{z}), \mathbf{z})$. Fourthly, $G_D(G_A(G_B(\mathbf{o}, \mathbf{z}'), \mathbf{z}), \mathbf{z})$ is translated back to domain O using G_C, which outputs $G_C(G_D(G_A(G_B(\mathbf{o}, \mathbf{z}'), \mathbf{z}), \mathbf{z}), \mathbf{z}')$ as the reconstructed version of $\mathbf{o}$.

Discriminators. The discriminator D_A is learned by discriminating between the real transition signal $\mathbf{o}$ of domain O and the fake outputs of G_A, while the discriminator D_B is learned by discriminating between the real EEG signal $\mathbf{u}$ of domain U and the fake outputs of G_B. Similarly, the discriminator D_D is learned by discriminating between the real speech signal $\mathbf{v}$ of domain V and the fake outputs of G_D, while the discriminator D_C is learned by discriminating between the generated transition signal $G_A(G_B(\mathbf{o}, \mathbf{z}'), \mathbf{z})$ of domain O and the fake outputs of G_C.

The generators G_A, G_B, G_C and G_D are optimized to emulate the fake outputs to fool the corresponding discriminators D_A, D_B, D_C and D_D as well as to minimize the following reconstruction losses $\|\mathbf{u} - G_B(G_A(\mathbf{u}, \mathbf{z}), \mathbf{z}')\|$, $\|\mathbf{o} - G_A(G_B(\mathbf{o}, \mathbf{z}'), \mathbf{z})\|$, $\|\mathbf{v} - G_D(G_C(\mathbf{v}, \mathbf{z}'), \mathbf{z})\|$, and $\|G_A(G_B(\mathbf{o}, \mathbf{z}'), \mathbf{z}) - G_C(G_D(G_A(G_B(\mathbf{o}, \mathbf{z}'), \mathbf{z}), \mathbf{z}), \mathbf{z}')\|$.

Loss functions. The same loss function is used for generators G_A and G_B as they perform the same task [38], which is defined as

$$\begin{aligned}\mathcal{L}(U, O)^G &= \lambda_U \|\mathbf{u} - G_B(G_A(\mathbf{u}, \mathbf{z}), \mathbf{z}')\| \\ &+ \lambda_O \|\mathbf{o} - G_A(G_B(\mathbf{o}, \mathbf{z}'), \mathbf{z})\| \\ &- D_A(G_A(\mathbf{u}, \mathbf{z})) - D_B(G_B(\mathbf{o}, \mathbf{z}')),\end{aligned} \tag{4}$$

where λ_U and λ_O are two constant parameters, which are typically set to the values within [100, 1000] [38].

Analogously, the loss function for both generators G_C and G_D is defined as follows

$$\begin{aligned}&\mathcal{L}(V,O)^G = \lambda_V \|\mathbf{v} - G_D(G_C(\mathbf{v},\mathbf{z}'),\mathbf{z})\| \\ &+ \lambda_O \|G_A(G_B(\mathbf{o},\mathbf{z}'),\mathbf{z}) - G_C(G_D(G_A(G_B(\mathbf{o},\mathbf{z}'),\mathbf{z}),\mathbf{z}),\mathbf{z}')\| \\ &- D_D(G_D(G_A(G_B(\mathbf{o},\mathbf{z}'),\mathbf{z}),\mathbf{z})) - D_C(G_C(\mathbf{v},\mathbf{z}')),\end{aligned} \tag{5}$$

where λ_V is a constant parameter which can be set typically to the values within [100, 1000].

The loss functions of D_A, D_B, D_C and D_D advocated by Wasserstein GAN (WGAN) [38] can be described by

$$\mathcal{L}_A^D = D_A(G_A(\mathbf{u},\mathbf{z})) - D_A(\mathbf{o}), \tag{6}$$

$$\mathcal{L}_B^D = D_B(G_B(\mathbf{o},\mathbf{z}')) - D_B(\mathbf{u}), \tag{7}$$

$$\mathcal{L}_C^D = D_C(G_C(\mathbf{v},\mathbf{z}')) - D_C(G_A(G_B(\mathbf{o},\mathbf{z}'),\mathbf{z})), \tag{8}$$

$$\mathcal{L}_D^D = D_D(G_D(G_A(G_B(\mathbf{o},\mathbf{z}'),\mathbf{z}),\mathbf{z})) - D_D(\mathbf{v}), \tag{9}$$

where $D_A(\cdot)$, $D_B(\cdot)$, $D_C(\cdot)$ and $D_D(\cdot)$ are defined similarly as in (2) and (3).

With adversarial training of the proposed Dual-DualGAN, we find the mapping between the EEG signals $\mathbf{u} \in U$ and the transition signals $\mathbf{o} \in O$, and the mapping between the speech signals $\mathbf{v} \in V$ and the transition signals $\mathbf{o} \in O$. Thus, the transition signals $\mathbf{o} \in O$ can be considered as shared labels for the EEG signals $\mathbf{u} \in U$ and the speech signals $\mathbf{v} \in V$ without corresponding features, which facilitates the one-to-one translation from EEG to speech.

5.2. Testing

As shown in Fig. 4(b), with adversarial training of the proposed Dual-DualGAN, we can obtain the trained parameters of the Dual-DualGAN. The EEG signals $\mathbf{u} \in U$ as the inputs can be translated to the speech signals $\mathbf{v} \in V$ by using the trained parameters of the Dual-DualGAN, which realize one-to-one EEG-to-speech translation.

5.3. Network configuration

Fig. 4(a) includes three domains: the domain U with the EEG signals $\mathbf{u}$, the domain V with the speech signals $\mathbf{v}$, and the transition domain O with the transition signals $\mathbf{o}$. By inserting the transition domain O with the transition signals $\mathbf{o}$, the cross-domain EEG-to-speech translation can be realized.

The transition signals $\mathbf{o}$ are obtained by cascading the corresponding EEG and speech signals in a certain proportion. As illustrated in Fig. 5, the EEG and speech signals are cascaded in different proportions from 1:4 to 4:1 with the step of $\frac{1}{5}$. For low values of the proportion 1:4 and 2:3, the accuracy rates are relatively low (at around 0.63 and 0.82, respectively). When the values of the proportion are higher than 1:1, the accuracy can be increased to above 0.88, with the highest value 0.95 achieved for the proportion of 3:2. Thus, we choose the proportion of 3:2 to cascade the EEG and speech signals in this paper.

The proposed network is summarized in Algorithm 1.

6. Data collection and preprocessing

6.1. Participants and speech datasets

Participants for data collection in the study were students and academic staff from Taiyuan University of Science and Technology, all in good health, including 24 male and 26 female, aged between 20 and 40. All participants washed their hair before the experiment to ensure their scalps were clean. In addition, they were not allowed to wear any jewelry. In the experiments, the participants were asked to place their forearms and hands in a place where they feel comfortable without movements, and to relax as much as possible in order to reduce facial muscle movements and eye blinking.

Algorithm 1 Dual-DualGAN

```
Input: EEG signals $\mathbf{u} \in U$, speech signals $\mathbf{v} \in V$, transition signals
    $\mathbf{o} \in O$, the number of critic iterations per generator iteration $N$,
    $\lambda_U$, $\lambda_U$, $\lambda_O$, an initial learning rate, and batch size $M$, which are
    depicted in Section 5.
Output: One-to-one EEG-to-speech translation of $\mathbf{u} \in U$ to $\mathbf{v} \in V$.
  1. Loss function of generators $G_A$ and $G_B$ in DualGAN 1.
  $\mathcal{L}(U,O)^G = \lambda_U \|\mathbf{u} - G_B(G_A(\mathbf{u},\mathbf{z}),\mathbf{z}')\|$
  $+\lambda_O \|\mathbf{o} - G_A(G_B(\mathbf{o},\mathbf{z}'),\mathbf{z})\|$
  $-D_A(G_A(\mathbf{u},\mathbf{z})) - D_B(G_B(\mathbf{o},\mathbf{z}'))$,
  where $\lambda_U$ and $\lambda_O$ are defined as in (4).
  2. Loss function of generators $G_C$ and $G_D$ in DualGAN 2.
  $\mathcal{L}(V,O)^G = \lambda_V \|\mathbf{v} - G_D(G_C(\mathbf{v},\mathbf{z}'),\mathbf{z})\|$
  $+\lambda_O \|G_A(G_B(\mathbf{o},\mathbf{z}'),\mathbf{z})$
  $-G_C(G_D(G_A(G_B(\mathbf{o},\mathbf{z}'),\mathbf{z}),\mathbf{z}),\mathbf{z}')\|$
  $-D_D(G_D(G_A(G_B(\mathbf{o},\mathbf{z}'),\mathbf{z}),\mathbf{z})) - D_C(G_C(\mathbf{v},\mathbf{z}'))$,
  where $\lambda_V$ is defined as in (5).
  3. Loss function of discriminators $D_A$, $D_B$, $D_C$ and $D_D$.
  $\mathcal{L}_A^D = D_A(G_A(\mathbf{u},\mathbf{z})) - D_A(\mathbf{o})$,
  $\mathcal{L}_B^D = D_B(G_B(\mathbf{o},\mathbf{z}')) - D_B(\mathbf{u})$,
  $\mathcal{L}_C^D = D_C(G_C(\mathbf{v},\mathbf{z}')) - D_C(G_A(G_B(\mathbf{o},\mathbf{z}'),\mathbf{z}))$,
  $\mathcal{L}_D^D = D_D(G_D(G_A(G_B(\mathbf{o},\mathbf{z}'),\mathbf{z}),\mathbf{z})) - D_D(\mathbf{v})$,
  where $D_A(\cdot)$, $D_B(\cdot)$, $D_C(\cdot)$ and $D_D(\cdot)$ are defined similarly as in (2)
  and (3).
  With adversarial training of the proposed Dual-DualGAN, the
  Dual-DualGAN learns the mapping between $\mathbf{u} \in U$ and $\mathbf{v} \in V$.
```

The non-invasive EEG signals measuring human neural activity were collected as the participants listen to continuous speech audio with a dedicated earphone. The speech signals were taken from the TIMIT[2] dataset which contains 6300 sentences, spoken by 630 speakers (438 male and 192 female, sampled at 16 kHz). We consider the sentences of the above dataset for training and testing.

6.2. Experimental paradigm with supervision

To improve the efficiency of auditory speech stimuli, an experimental paradigm with supervision based on the traditional experimental paradigm is proposed by considering participant's attention detected with a threshold, as illustrated in Fig. 6(a). The temporal events in each experiment for data capture are shown in Fig. 6(b).

To ensure the quality of the EEG signals recorded, we design a device for measuring the attention of participants in response to the stimuli played using the TGAM (ThinkGear$^{\text{TM}}$ Asic Module) produced by NeuroSky (see Fig. 7). We start recording the EEG signal only when the attention is higher than a pre-defined threshold. The attention P can be described as follows based on the eSense$^{\text{TM}}$ algorithm.

$$P = \frac{l\beta + m\beta}{\theta}, \tag{10}$$

where $m\beta$ is middle beta waves (frequency 16–20 Hz), $l\beta$ is low beta waves (frequency 12–15 Hz), and θ is theta waves

[2] https://catalog.ldc.upenn.edu/docs/LDC93S1/TIMIT.html.

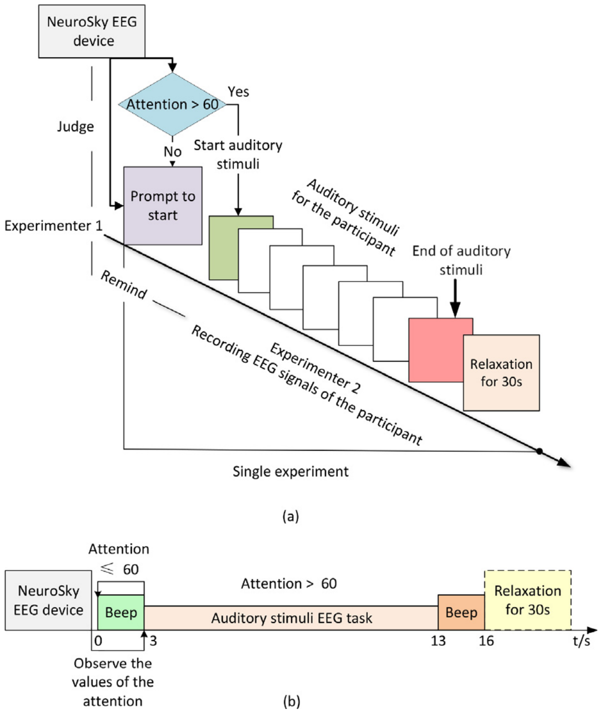


Fig. 6. Design of the experimental paradigm for EEG data collection, (a) experimental paradigm with supervision, and (b) temporal events in each experiment.

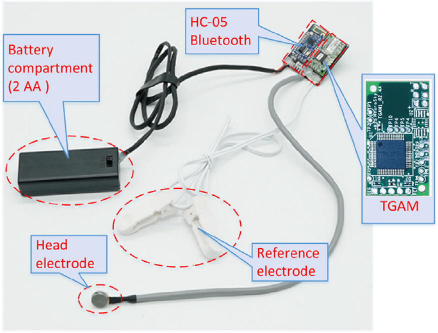


Fig. 7. Device for detecting participants' attention.

(frequency 4–7 Hz). The attention with a value greater than a predefined threshold e.g. $P > 60$ indicates that the participants are concentrating on the auditory stimuli, and the EEG signals can be recorded from this moment.

In Fig. 6, at the beginning of an experiment, the experimenter 1 plays a beep to remind the participant concentrating on the experiment, and observes the participant's attention P for 3 s. For $P \leq 60$, the experimenter 1 repeats the above procedure. For $P > 60$, the experimenter 1 starts to play a continuous speech file and reminds the experimenter 2 recording the EEG signals, and each speech file is repeated at least five times. At the end of the experiment, the experimenter 1 plays a beep to remind the experimenter 2 stopping the recording, and the participants can open their eyes, blink and relax. After relaxing for 30 s, they can start the next experiment.

6.3. Data collection and EEG datasets

In the experiments, a data collection platform is set up for the non-invasive EEG neural recordings. The EEG signals are recorded from 24 electrodes placed around the scalp according to international 10–20 system by using Electroencephalogram and evoked potentiometer NCERP produced by Shanghai Nuocheng Electric Co., Ltd (NCC). By connecting the electrode cap to the physiological amplifier, analog EEG signals are collected and amplified. By using the optical fibers for transmitting the data to the EEG master control box, the amplified EEG signals are digitized at about 8 kHz and 32 bit, filtered with the cut off frequencies of 1 Hz and 50 Hz, the channels with visible artifact or excessive noise are removed, and then transmitted to the computer by USB interface. The platform for the collection of EEG signals is shown in Fig. 8, and the parameters of NCERP are listed in Table 1.

We choose four-channel EEG signals according to the EEG electrode position of the temporal lobe (see Fig. 9 with red boxes).

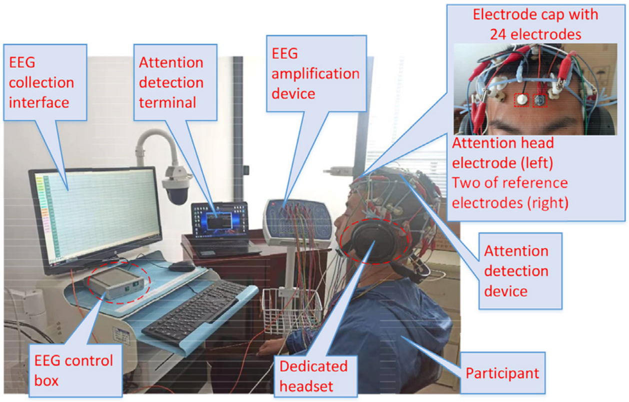

EEG collection interface
Attention detection terminal
EEG amplification device
Electrode cap with 24 electrodes
Attention head electrode (left)
Two of reference electrodes (right)
Attention detection device
EEG control box
Dedicated headset
Participant

Fig. 8. Data collection platform for EEG signals.

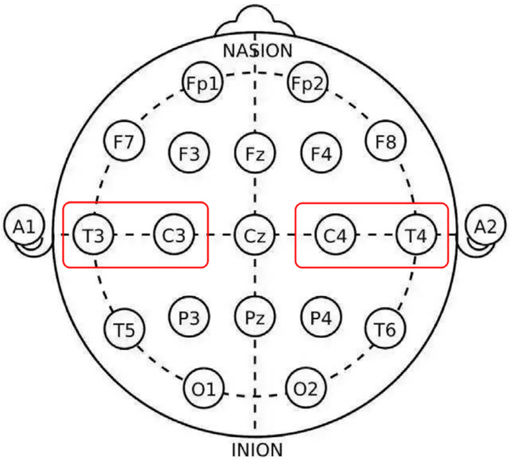


Fig. 9. Placement of twenty-four EEG electrodes, and four of which are selected according to the temporal lobe and highlighted with red boxes.

Table 1
The parameters of NCERP.

Parameter	Value
Calibration voltage	100 μV, error ≤ ±5%
Sensitivity	5 μV/cm, error ≤ ±5%
Time constants	0.1 s, 0.2 s, 0.3 s, and error ≤ ±10%
Noise level	<0.3 μV (RMS)
Rejection ratio	≥110 dB
Amplitude frequency characteristic	1 Hz∼60 Hz, and error +3% ∼−15%
Polarization resistance voltage	±300 mV DC polarization voltage, sensitive change ±5%
Input impedance	≥10 MΩ

The selected EEG signals are normalized and transformed into a multi-dimensional matrix to build a new EEG dataset for the participants with good attention, which will be the input of the proposed Dual-DualGAN.

7. Experimental setup

7.1. Setup for the proposed method

In our Dual-DualGAN, the generators are the U-shaped net as in [46] configured with equal number of down-sampling (pooling) and up-sampling layers. The down-sampling (pooling) layers are constructed by eight convolution layers with the kernel of (3, 3), each neuron with a LeakyReLU activation function, the step of convolution is 1, and the step of pooling is 2. The up-sampling layers are constructed by eight deconvolution layers with the kernel of (3, 3), each neuron with a LeakyReLU activation function, and the step of convolution is 1. The skip connections between mirrored down-sampling and up-sampling layers are used to enable the low-level information to be shared between the input and output and to avoid loss of information. The discriminators are the Markovian Patch-GAN as in [47], which are constructed by five convolution layers with the kernel of (3, 3), and has no constraints over the size of the input signal. The number of critic iterations per generator iteration N can be set to 5, λ_U, λ_U and λ_O are all set to 500, an initial learning rate is set at 0.0002, and the batch size M is assigned with 1.

7.2. Baseline methods

We compare several versions of our Dual-DualGAN algorithm: M1 (removing two convolution layers of the generative network), M2 (increasing two convolution layers of the generative network), M3 (removing two convolution layers of the discriminative network), M4 (increasing two convolution layers of the discriminative network), M5 (setting the step of convolution to 2), M6 (setting the convolution layers with the kernel of (5, 5)), M7 (changing activation function to ReLU), Dual-DualGAN (full version of our Dual-DualGAN algorithm) with the state-of-the-art algorithms: Deep Neural Network (DNN)-based encoder–decoder algorithm[3] [16], bidirectional Long and Short Term Memory Network (bLSTM)-based encoder–decoder algorithm[4] [17], Recurrent Neural Network (RNN)-based encoder–decoder algorithm [34], and DualGAN[5] [38].

All the algorithms are trained from scratch by using cross validation, by randomly picking 80% data of the EEG dataset for training, the remaining 20% data for testing. The facilities used to perform the experiments include Intel I9-10900X 13.7 GHz CPU, 2*NVIDIA RTX 8000 Graphics Card and 6 $*$ 32 GB memory.

7.3. Performance metrics

For performance evaluation, we use the accuracy rate [48], Pearson correlation coefficient (PCC) [49] and Mel-cepstral distortion (MCD) [50] as the performance metrics.

The accuracy metric is defined as the proportion of correct predictions among the total number of cases examined.

$$Accuracy = \frac{T}{T+F}, \tag{11}$$

where T means the correct predictions and F means the false predictions.

Table 2
The performance and time complexity of the Dual-DualGAN as compared with state-of-the-art algorithms.

Algorithm		Accuracy rate (%)	PCC	MCD (dB)
Encoder–decoder framework	DNN	74.87	0.771	4.154
	bLSTM	70.04	0.729	4.437
	RNN	71.10	0.747	4.352
DualGAN		56.82	0.624	5.015
M1		74.24	0.792	3.947
M2		77.92	0.829	3.808
M3		73.86	0.788	3.971
M4		78.18	0.834	3.798
M5		77.03	0.821	3.815
M6		77.48	0.828	3.811
M7		75.83	0.796	3.941
Dual-DualGAN		**78.53**	**0.838**	**3.793**

The PCC is a measure of linear correlation between the original and the synthesized speech signal, defined as

$$PCC = \frac{\mathrm{cov}(\mathbf{v}\mathbf{v}')}{\sigma_{\mathbf{v}}\sigma_{\mathbf{v}'}}, \tag{12}$$

where $\mathrm{cov}(\mathbf{v}\mathbf{v}')$ is the covariance of the original speech signal $\mathbf{v}$ and the synthesized speech signal $\mathbf{v}'$, and $\sigma_{\mathbf{v}}$ and $\sigma_{\mathbf{v}'}$ are the standard deviation of $\mathbf{v}$ and $\mathbf{v}'$, respectively.

The metric $MCD(k)$ evaluates objective speech quality, defined as

$$MCD(k) = \frac{10}{\ln 10}\frac{1}{T}\sum_{i=0}^{T-1}\sqrt{\sum_{k=1}^{K}(mc(i,k) - mc'(i,k))^2}, \tag{13}$$

where $mc(i,k)$ and $mc'(i,k)$ are the ith mel-cepstral coefficient of the kth frame of the original and the synthesized speech signal, respectively.

8. Results

In this section, we carry out experiments to demonstrate the performance of the proposed Dual-DualGAN, and how the algorithm is affected for solving the ET-CAS problem.

To evaluate the performance of our Dual-DualGAN, we conduct three listening tasks that involve word-level, short-sentence-level (no more than six words) and long-sentence-level (more than six words) transcription, respectively. The word-level speech signals are separated from the sentence-level speech signals.

In Table 2, we compare several versions of our Dual-DualGAN algorithm with the state-of-the-art algorithms. The results show that the proposed Dual-DualGAN has better performance in average accuracy rate, PCC and MCD than the RNN-based, bLSTM-based, DNN-based encoder–decoder algorithms, and DualGAN. The accuracy rates and PCCs of the RNN-based, bLSTM-based, DNN-based encoder–decoder algorithms, and DualGAN are less than 74.9% and 0.78, and the MCDs of the RNN-based, bLSTM-based, DNN-based encoder–decoder algorithms, and DualGAN are more than 4.1 dB.

Fig. 10 shows the spectrograms of the speech signals from the original word, short sentence and long sentence, respectively, and those decoded from human neural activity. All the synthesized spectrograms retain salient energy patterns that are present in the original spectrograms.

As listed in Table 3, we found that for word-level translation, if the mapping is correct, the accuracy can reach 100%. For the short and long sentences, which are composed of several words and are relatively complex, there is a decrease in the translation accuracy due to the incorrect mapping of some words. Therefore, for some

[3] http://naplab.ee.columbia.edu/naplib.html.

[4] https://doi.org/10.1038/s41586-019-1119-1.

[5] https://github.com/duxingren14/DualGAN.

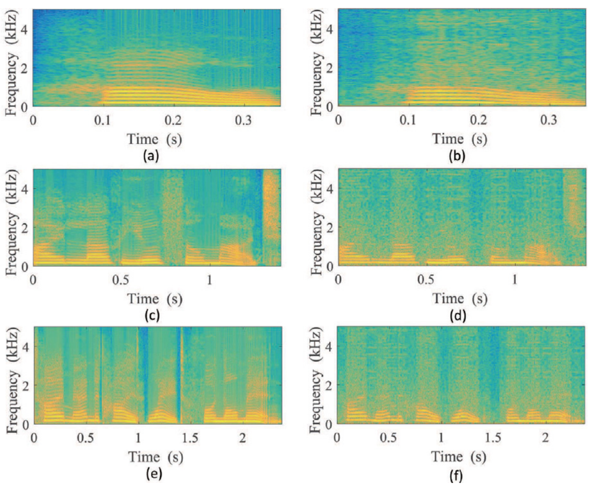


Fig. 10. Spectrograms of the synthesized speech and the original speech signals, (a) the original spectrogram of the speech signal for a word, (b) the synthesized spectrogram of the speech signal for a word, (c) the original spectrogram of the speech signal for a short sentence, (d) the synthesized spectrogram of the speech signal for a short sentence, (e) the original spectrogram of the speech signal for a long sentence, and (f) the synthesized spectrogram of the speech signal for a long sentence.

Table 3
Listener transcriptions of neurally synthesized speech for different words and sentences.

Type	Accuracy rate	Original words or sentences and transcriptions of synthesized speech
Word	100%	O: weekend T: weekend
Short sentence	80%	O: I love you so much. T: I laugh you so much.
	60%	O: Eat more fish, less meat. T: Beat more fish, less meet.
Long sentence	75%	O: A good beginning makes for a good ending. T: A good beginning made for a great ending.
	60%	O: What makes the desert beautiful is that somewhere it hides a well. T: What made the desert beautiful was that flower it his a dog.

sentences, the accuracy drops to less than 80%, as their similarity to the original sentences has reduced. However, the synthesized spectrograms for the short and long sentences retain most of the salient energy patterns of the original spectrograms.

In Fig. 11(a), we compare the average accuracy rates of the synthesized speech signals for word-level, short-sentence-level and long-sentence-level transcription. The average accuracy rate of the word-level transcription is higher than those of the short-sentence-level and long-sentence-level transcription, and all the values of the accuracy rate are around 78.5%. The average PCCs between the synthesized speech signals and the original speech signals are shown in Fig. 11(b). The PCCs of the word-level, short-sentence-level and long-sentence-level transcription are relatively high and the values are above 0.83. We also compare the average MCD of the synthesized speech signals (see Fig. 11(c)). The MCDs of the synthesized speech signals for word-level, short-sentence-level and long-sentence-level transcription are relatively small, and the values of MCDs are about 3.9. This demonstrates the efficiency of the proposed Dual-DualGAN in decoding speech from human neural activity.

The gender effects are considered on listener transcriptions of the neurally synthesized speech. The data of 20 male and 20 female are randomly selected from the EEG datasets. The average accuracy rates of the synthesized speech signals for the word-level, short-sentence-level and long-sentence-level transcription by gender are illustrated in Fig. 12. The average accuracy rates of male for word-level, short-sentence-level and long-sentence-level transcription are 78.5%, 78.3%, and 77.9%, and the average accuracy rates of female are 79.1%, 78.9%, and 78.6%. It shows the efficiency and adaptability of the proposed Dual-DualGAN, regardless of gender.

The age effects are also considered on listener transcriptions of the neurally synthesized speech. The data of four age groups (20–25, 25–30, 30–35, and 35–40 years old) of the participants are randomly picked from the EEG datasets. The average accuracy rates of the synthesized speech signals for word-level, short-sentence-level and long-sentence-level transcription for these age groups are illustrated in Fig. 13. The average accuracy rates for word-level, short-sentence-level and long-sentence-level transcription of the participants in age between 25 to 30 and 30 to 35 years old are mostly above 78.5%. The average accuracy rates of the participants in age between 20 to 25 and 35 to 40 years old are slightly low, and the values are about 78%. The results demonstrate the proposed Dual-DualGAN can translate an EEG to a speech signal with a good generalization ability for different age groups.

9. Conclusion

We have presented a new method for the problem of end-to-end translation from human neural activity to speech (ET-CAS). Our contributions to this challenging problem are as follows:

Model. We have formulated an end-to-end model for the ET-CAS problem, i.e. translating human neural activity to speech directly.

Datasets. We developed a new EEG dataset where the attention of the participants is detected and used to guide the collection of the EEG signals in each experiment.

Network. We proposed a dual–dual generative adversarial network (Dual-DualGAN) to address the ET-CAS problem. In this system, two DualGANs are created and trained simultaneously, where a transition domain is introduced into the DualGAN to bridge the two DualGANs. The EEG and speech signals are cascaded proportionally to generate the transition signals i.e. constructing shared labels for EEG and speech signals without mapping their features.

Numerical experiments show that the proposed ET-CAS algorithm performs well in translating human neural activity to speech. In the future, it is interesting to investigate how to incorporate acoustic and emotional features into the ET-CAS model and algorithm, which may improve the performance of the system in decoding speech signals that consist of sentences with repetitive words.

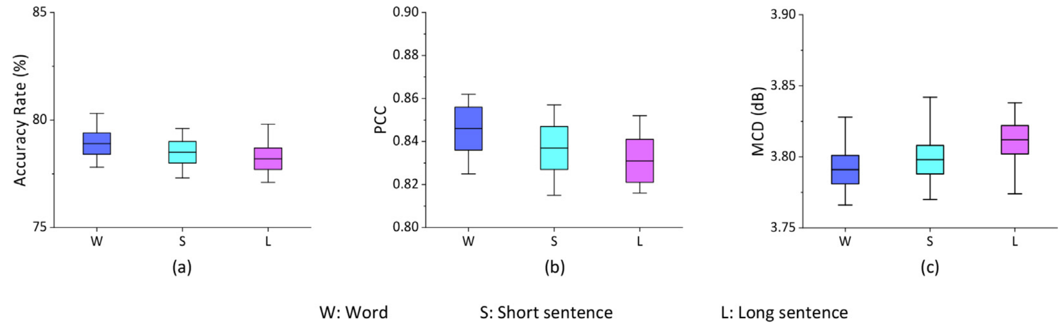


Fig. 11. Performance evaluations of the neurally synthesized speech for the word-level, short-sentence-level and long-sentence-level transcription, (a) average accuracy rate of the synthesized speech signals, (b) average PCC between the synthesized speech signals and the original speech signals, and (c) average MCD of the synthesized speech signals in comparison with the original speech signals.

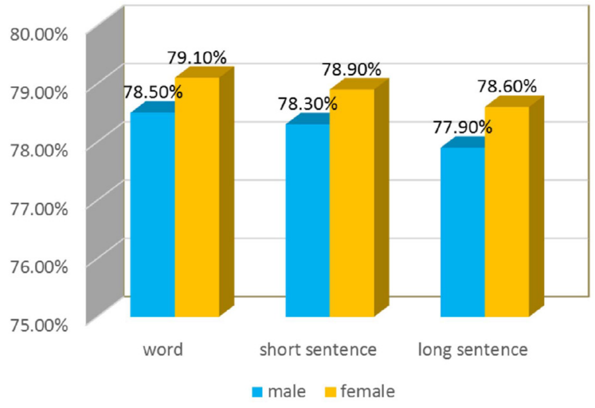


Fig. 12. Average accuracy rates of the synthesized speech signals for word-level, short-sentence-level and long-sentence-level transcription by gender.

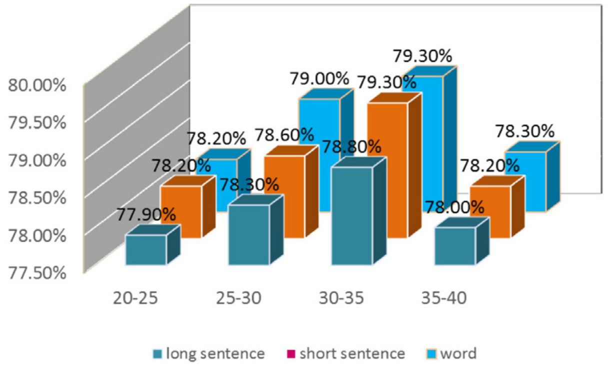


Fig. 13. Average accuracy rates of the synthesized speech signals for word-level, short-sentence-level and long-sentence-level transcription by age.

CRediT authorship contribution statement

Yina Guo: Conceptualization, Methodology, Writing – original draft. **Ting Liu:** Resources, Validation. **Xiaofei Zhang:** Data curation, Resources, Validation. **Anhong Wang:** Supervision. **Wenwu Wang:** Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Yina Guo reports financial support was provided by Taiyuan University of Science and Technology. Yina Guo reports a relationship with National Natural Science Foundation of China, Science and Technology Innovation Talent Team of Shanxi Province, Shanxi Province Postgraduate Excellent Innovation Project Plan, Shanxi Scholarship Council of China that includes: funding grants.

Data availability

The data that support the findings of this study are available https://github.com/qwe1218088/dual-dualgan.git.

Acknowledgments

The authors would like to thank the associate editor and anonymous reviewers for their constructive comments for improving this paper. This work was supported by National Natural Science Foundation of China under Grant 62271341, Science and Technology Innovation Talent Team of Shanxi Province under Grant 202204051001018, Shanxi Province Postgraduate Excellent Innovation Project Plan under Grant 2021Y679 and 2022Y689, Shanxi Scholarship Council of China under Grant HGKY2019080 and 2020-127. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

Code availability

The codes for reproducing the results are available at https://github.com/qwe1218088/dual-dualgan.git. The codes can be used only for non-commercial purpose.

References

[1] C. Du, K. Fu, J. Li, H. He, Decoding visual neural representations by multimodal learning of brain-visual-linguistic features, IEEE Trans. Pattern Anal. Mach. Intell. (2023) 1–17.

[2] P. Singh, P. Pandey, K. Miyapuram, S. Raman, EEG2image: Image reconstruction from EEG brain signals, in: 2023 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2023, pp. 1–5.

[3] R. Manor, L. Mishali, A.B. Geva, Multimodal neural network for rapid serial visual presentation brain computer interface, Front. Comput. Neurosci. 10 (2016) 130.

[4] J. Jin, H. Zhang, I. Daly, X. Wang, A. Cichocki, An improved P300 pattern in BCI to catch user's attention, J. Neural Eng. 14 (3) (2017) 036001.

[5] J.J. Norton, S. Umunna, T. Bretl, The elicitation of steady-state visual evoked potentials during sleep, Psychophysiology 54 (4) (2017) 496–507.

[6] M. Guo, J. Jin, Y. Jiao, X. Wang, A. Cichockia, Investigation of visual stimulus with various colors and the layout for the oddball paradigm in evoked related potential-based brain–computer interface, Front. Comput. Neurosci. 13 (2019) 24.

[7] J.D. Chailloux Peguero, O. Mendoza-Montoya, J.M. Antelis, Single-option P300-BCI performance is affected by visual stimulation conditions, Sensors-Basel 20 (24) (2020) 7198.

[8] P.R. Bassi, W. Rampazzo, R. Attux, Transfer learning and SpecAugment applied to SSVEP based BCI classification, Biomed. Signal Process. 67 (2021) 102542.

[9] F. Nijboer, A. Furdea, I. Gunst, J. Mellinger, D.J. McFarland, N. Birbaumer, A. Kübler, An auditory brain–computer interface (BCI), J. Neurosci. Methods 167 (1) (2008) 43–50.

[10] D.S. Klobassa, T.M. Vaughan, P. Brunner, N. Schwartz, J.R. Wolpaw, C. Neuper, E. Sellers, Toward a high-throughput auditory P300-based brain–computer interface, Clin. Neurophysiol. 120 (7) (2009) 1252–1261.

[11] A. Kübler, A. Furdea, S. Halder, E.M. Hammer, F. Nijboer, B. Kotchoubey, A brain–computer interface controlled auditory event-related potential (P300) spelling system for locked-in patients, Ann. NY Acad. Sci. 1157 (1) (2009) 90–100.

[12] K.-W. Lee, D.-H. Lee, S.-J. Kim, S.-W. Lee, Decoding neural correlation of language-specific imagined speech using EEG signals, in: 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society, EMBC, 2022, pp. 1977–1980.

[13] J. Höhne, M. Schreuder, B. Blankertz, M. Tangermann, A novel 9-class auditory ERP paradigm driving a predictive text entry system, Front. Neurosci-Switz. 5 (2011) 99.

[14] J. Höhne, K. Krenzlin, S. Dähne, M. Tangermann, Natural stimuli improve auditory BCIs with respect to ergonomics and performance, J. Neural Eng. 9 (4) (2012) 045003.

[15] F. Bocquelet, T. Hueber, L. Girin, C. Savariaux, B. Yvert, Real-time control of an articulatory-based speech synthesizer for brain computer interfaces, PLoS Comput. Biol. 12 (11) (2016) e1005119.

[16] H. Akbari, B. Khalighinejad, J.L. Herrero, A.D. Mehta, N. Mesgarani, Towards reconstructing intelligible speech from the human auditory cortex, Sci. Rep-UK 9 (1) (2019) 1–12.

[17] G.K. Anumanchipalli, J. Chartier, E.F. Chang, Speech synthesis from neural decoding of spoken sentences, Nature 568 (7753) (2019) 493–498.

[18] Z. Wang, H. Ji, Open vocabulary electroencephalography-to-text decoding and zero-shot sentiment classification, in: AAAI Conference on Artificial Intelligence, vol. 36, 2021, pp. 5350–5358.

[19] J.G. Makin, D.A. Moses, E.F. Chang, Machine translation of cortical activity to text with an encoder–decoder framework, Nature Neurosci. 23 (4) (2020) 575–582.

[20] F.R. Willett, D.T. Avansino, L.R. Hochberg, J.M. Henderson, K.V. Shenoy, High-performance brain-to-text communication via imagined handwriting, BioRxiv (2020).

[21] D.-W. Kim, H.-J. Hwang, J.-H. Lim, Y.-H. Lee, K.-Y. Jung, C.-H. Im, Classification of selective attention to auditory stimuli: Toward vision-free brain–computer interfacing, J. Neurosci. Methods 197 (1) (2011) 180–185.

[22] A. De Vos, R. Luke, H. Poelmans, M. Hofmann, J. Vanderauwera, M. Vandermosten, P. Ghesquière, J. Wouters, Lateralization of auditory steady-state responses to speech envelope modulations, in: International Evoked Response Audiometry Study Group (IERASG), Date: 2013/06/10-2013/06/13, Location: New Orleans, Louisiana, USA, 2013.

[23] K. Joos, A. Gilles, P. Van de Heyning, D. De Ridder, S. Vanneste, From sensation to percept: the neural signature of auditory event-related potentials, Neurosci. Biobehav. R. 42 (2014) 148–156.

[24] S. Halder, I. Käthner, A. Kübler, Training leads to increased auditory brain–computer interface performance of end-users with motor impairments, Clin. Neurophysiol. 127 (2) (2016) 1288–1296.

[25] J. Heo, H.J. Baek, S. Hong, M.H. Chang, J.S. Lee, K.S. Park, Music and natural sounds in an auditory steady-state response based brain–computer interface to increase user acceptance, Comput. Biol. Med. 84 (2017) 45–52.

[26] D. Hübner, A. Schall, N. Prange, M. Tangermann, Eyes-closed increases the usability of brain-computer interfaces based on auditory event-related potentials, Front. Hum. Neurosci. 12 (2018) 391.

[27] M. Huang, J. Jin, Y. Zhang, D. Hu, X. Wang, Usage of drip drops as stimuli in an auditory P300 BCI paradigm, Cogn. Neurodyn. 12 (1) (2018) 85–94.

[28] H. Akbari, B. Khalighinejad, J. Herrero, A. Mehta, N. Mesgarani, Towards reconstructing intelligible speech from the human auditory cortex, Sci. Rep. 9 (1) (2019) 1–12.

[29] G.K. Anumanchipalli, J. Chartier, E.F. Chang, Speech synthesis from neural decoding of spoken sentences, Nature 568 (2019) 493–498.

[30] A. Défossez, C. Caucheteux, J. Rapin, O. Kabeli, J.-R. King, Decoding speech from non-invasive brain recordings, 2022, arXiv preprint arXiv:2208.12266v1.

[31] J. Zhou, Y. Duan, Y. Zou, Y.-C. Chang, Y.-K. Wang, C.-T. Lin, Speech2EEG: Leveraging pretrained speech model for EEG signal recognition, IEEE Trans. Neural Syst. Rehabil. Eng. 31 (2023) 2140–2153.

[32] F. Cui, L. Guo, L. He, J. Liu, E. Pei, Y. Wang, D. Jiang, Relate auditory speech to eeg by shallow-deep attention-based network, in: 2023 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2023, pp. 1–2.

[33] N. Das, J. Vanthornhout, T. Francart, A. Bertrand, Stimulus-aware spatial filtering for single-trial neural response and temporal response function estimation in high-density EEG with applications in auditory research, Neuroimage 204 (2020) 116211.

[34] G. Krishna, C. Tran, Y. Han, M. Carnahan, A.H. Tewfik, Speech synthesis using EEG, in: 2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2020, pp. 1235–1238.

[35] F. Velasco-Álvarez, Á. Fernández-Rodríguez, M.T. Medina-Juliá, R. Ron-Angevin, Speech stream segregation to control an ERP-based auditory BCI, J. Neural Eng. 18 (2) (2021) 026023.

[36] D. Bank, N. Koenigstein, R. Giryes, Autoencoders, 2020, arXiv preprint arXiv:2003.05991.

[37] M. Sewak, S.K. Sahay, H. Rathore, An overview of deep learning architecture of deep neural networks and autoencoders, J. Comput. Theor. Nanosci. 17 (1) (2020) 182–188.

[38] Z. Yi, H. Zhang, P. Tan, M. Gong, DualGAN: unsupervised dual learning for image-to-image translation, in: Proceedings of the IEEE International Conference on Computer Vision, ICCV, 2017, pp. 2849–2857.

[39] J. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2223–2232.

[40] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, Adv. Neural Inf. Process. Syst. 33 (2020) 6840–6851.

[41] C. Herff, D. Heger, A. De Pesters, D. Telaar, P. Brunner, G. Schalk, T. Schultz, Brain-to-text: decoding spoken phrases from phone representations in the brain, Front. Neurosci-Switz. 9 (2015) 217.

[42] J.S. Brumberg, P.R. Kennedy, F.H. Guenther, Artificial speech synthesizer control by brain-computer interface, in: Proceedings of the Annual Conference of the International Speech Communication Association, 2009, pp. 636–639.

[43] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, Adv. Neural Inform. Proces. Syst. 3 (2014) 2672–2680.

[44] S. Wen, W. Liu, Y. Yang, T. Huang, Z. Zeng, Generating realistic videos from keyframes with concatenated GANs, IEEE Trans. Circuits Syst. Video Technol. 29 (8) (2019) 2337–2348.

[45] Y. Cao, N. Liu, C. Zhang, T. Zhang, Z.-F. Luo, Synchronization of multiple reaction–diffusion memristive neural networks with known or unknown parameters and switching topologies, Knowl.-Based Syst. 254 (2022) 109595.

[46] Z. Zhou, M.M.R. Siddiquee, N. Tajbakhsh, J. Liang, UNet++: a nested U-Net architecture for medical image segmentation, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Springer, 2018, pp. 3–11.

[47] C. Li, M. Wand, Precomputed real-time texture synthesis with markovian generative adversarial networks, in: European Conference on Computer Vision, ECCV, Springer, 2016, pp. 702–716.

[48] M. Story, R.G. Congalton, Accuracy assessment: a user's perspective, Photogramm. Eng. Remote Sens. 52 (3) (1986) 397–399.

[49] J. Benesty, J. Chen, Y. Huang, I. Cohen, Pearson correlation coefficient, in: Noise Reduction in Speech Processing, Springer, 2009, pp. 1–4.

[50] R. Kubichek, Mel-cepstral distance measure for objective speech quality assessment, in: Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing, vol. 1, 1993, pp. 125–128, http://dx.doi.org/10.1109/PACRIM.1993.407206.