electronics

Article

D2StarGAN: A Near-Far End Noise Adaptive StarGAN for Speech Intelligibility Enhancement

Dengshi Li *, Chenyi Zhu and Lanxin Zhao

School of Artificial Intelligence, Jianghan University, Wuhan 430050, China
* Correspondence: reallds@jhun.edu.cn

Abstract: When using mobile communication, the voice output from the device is already relatively clear, but in a noisy environment, it is difficult for the listener to obtain the information expressed by the speaker with clarity. Consequently, speech intelligibility enhancement technology has emerged to help alleviate this problem. Speech intelligibility enhancement (IENH) is a technique that enhances speech intelligibility during the reception phase. Previous research has focused on IENH through normal versus different levels of Lombardic speech conversion, inspired by a well-known acoustic mechanism called the Lombard effect. However, these methods often lead to speech distortion and impair the overall speech quality. To address the speech quality degradation problem, we propose an improved (StarGAN)-based IENH framework by combining StarGAN networks with the dual discriminator idea to construct the conversion framework. This approach offers two main advantages: (1) Addition of a speech metric discriminator on top of StarGAN to optimize multiple intelligibility and quality-related metrics simultaneously; (2) a framework that is adaptive to different distal and proximal noise levels with different noise types. Experimental results from objective experiments and subjective preference tests show that our approach outperforms the baseline approach, and these enable IENH to be more widely used.

Keywords: speech intelligibility enhancement; Lombard effect; StarGAN; speech conversion

Citation: Li, D.; Zhu, C.; Zhao, L. D2StarGAN: A Near-Far End Noise Adaptive StarGAN for Speech Intelligibility Enhancement. *Electronics* **2023**, *12*, 3620. https://doi.org/10.3390/electronics12173620

Academic Editor: Chiman Kwan

Received: 30 June 2023
Revised: 3 August 2023
Accepted: 4 August 2023
Published: 27 August 2023

1. Introduction

Owing to the rapid development of mobile communication and artificial intelligence technology, telephone conversation has become a common avenue of voice communication. People can also interact with various mobile devices anytime and anywhere, but this type of communication often occurs in noisy environments. These complex and variable noise environments bring great interference to communication. As shown in Figure 1, the interference of environmental noise in mobile communication mainly comes from two stages: the "talking stage" (in the far-end) and the "listening stage" (in the near-end) [1]. The noise in the "speaking phase" has been relatively well suppressed by the development of hardware and software for a long time, which is represented by speech enhancement (SE) [2,3]. However, the speech intelligibility enhancement (IENH) technology in the "listening stage" is relatively lagging behind. The problem where listeners have difficulty in obtaining information in a noisy environment has not yet been solved. Finding ways to improve speech intelligibility enhancement has become an urgent problem in speech dialogue.

In mobile communication, noise in both the "speaking phase" and the "listening phase" causes significant interference. In the "speaking phase", the microphone picks up the speaker's voice and also picks up the noise in the speaker's environment. The cell phone then encodes the voice signal to form a code stream, which is sent to the receiver by the communication channel. In the "listening stage", the cell phone receives the speech signal and plays it back to the listener, during which the noise of the environment is inevitably

heard. Speech intelligibility enhancement (IENH) is a technology that targets the near-end of the receiver, and is thus called near-end listening enhancement (NELE) technology [4].

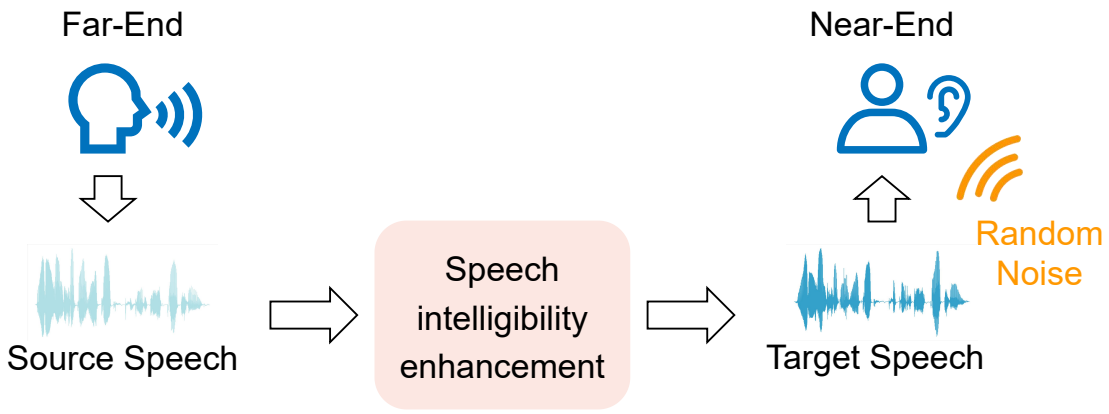

Figure 1. Diagram of telephone communication with environmental noise. Both source and target speech are clean speech with noise removed.

The perceptual efficiency of a speech signal is usually measured on the basis of speech intelligibility and speech quality. Speech intelligibility indicates the extent to which information can be understood by the listener and depends on the proportion of speech subunits that are correctly recognized (e.g., syllables, phonemes, phrases). Speech quality, on the other hand, is a multidimensional measure of the listener's subjective goodness or badness of a speech sample and encompasses the perceived quality, naturalness, and clarity of speech.

In real-time communication, the near-end listener is sometimes in an environment containing noise that can not be avoided and can not be avoided in advance of the near-end of the background noise, the near-end of the noise of the degree of noise, the noise is sharp, etc., which are difficult to control. Therefore, a solution considered by some is to enhance the source speech coming from the far end to reduce the effect of near-end noise on the reception of information, and to enable the near-end listener to receive effective information. Enhancement here consists of one of the simplest ways to increase the output power of the communication device, i.e., to increase the volume. However, there is an upper limit to the volume of a mobile communication device, and also excessive volume reduces the comfort of communication and may cause damage to the auditory nerves of the listener, tinnitus and other serious irreversible consequences.

Speech signals are produced by vibrations of the vocal cords, arising from a periodic excitation flow, and have many acoustic characteristics themselves, some of which have been shown to improve the robustness of speech information in noisy environments by changing these (e.g., Mel cepstrum, fundamental frequency, etc.). Therefore, without changing the output power of the mobile device, the acoustic features of the source speech signal can be changed through the IENH algorithm, so that the enhanced speech has higher intelligibility in the same noise environment. This plays a crucial role in improving the efficiency of speech communication.

Traditional SE methods are based on signal-processing methods, modeled spectral estimation methods [5–8]. Recently, deep learning (DL) models have become popular in the SE domain. Several network architectures, such as deep denoising autoencoders [9], fully connected networks [10], convolutional neural networks (CNNs) [11], recurrent neural networks (RNNs), and long-term short-term memory networks (LSTMs) [12], have demonstrated remarkable improvements in SE capabilities compared to traditional SE methods. With the development of speech enhancement (SE) technology [13,14], the noise in the remote voice signal has been suppressed to a greater extent during transmission to the listener through the device, and the audio played by the listener's cell phone is clear enough. However, it remains difficult to obtain the information from the audio played in

the cell phone clearly due to the influence of the surrounding environmental noise, so the attention of many scholars has slowly turned to speech intelligibility enhancement (IENH) technology.

In early IENH research, frequency domain energy redistribution and digital-signal-processing algorithms based on acoustic masking principles [15] were a common approach. However, this approach lacked the preservation of the naturalness of speech [16]. So researchers began to explore new data-driven methods. This new approach was inspired by the Lombard reflex [17,18] (or referred to as the Lombard effect) in human voice mechanisms. The Lombard effect suggests that a speaker's speaking style involuntarily changes in response to noise levels. The Lombard effect is also more pronounced when the noise level increases and the speaker's voice becomes clearer and more pronounced. And the change in spectral tilt is believed to be a key factor in the Lombard reflex to improve speech intelligibility. The data-driven approach uses a large amount of speech data and builds a model through deep learning algorithms, which enables the model to adaptively adjust the Melody Cepstrum of speech (MFCC) to achieve a more natural IENH. this approach has achieved good results in practical applications and has become an important research direction in the field of IENH [19].

A prerequisite for achieving accurate conversion of Lombard speech features to normal speech features is that the feature conversion model in the feature conversion framework is effective and efficient. Many previous studies (e.g., [19,20]) have relied on parallel datasets, relying on parallel speech pairs of source (normal) and target (Lombard) speech. However, under the Lombard effect, speakers usually speak at a slower rate. Parallel speech style conversion (SSC) requires a temporal alignment operation to pre-process the training data, which is performed by lossy algorithms (e.g., dynamic time warping (DTW)) and can lead to feature distortion. This encourages the use of non-parallel SCC to avoid time-aligned operations. In recent years, there has been much successful research in neural networks for mapping near-frequency cepstral coefficients (MFCCs). Some recent studies have combined variants of generative adversarial networks (GAN) [21–23], demonstrating promising results. These methods are used to train speech synthesizers without the need for parallel speech, transcribed or time-aligned data. They offer better intelligibility and naturalness than traditional methods.

However, these methods still face some limitations:

(1) In real life, the speaker cannot be in a completely noise-free environment, the far-end speaker may itself be speaking Lombard speech, and there are limitations to the method of simply converting normal speech to Lombard speech.
(2) Although the StarGAN-based approach is an improvement over the CycleGAN-based approach, the adequacy of feature mapping is still significantly deficient, and there remains an insurmountable gap between real speech and transformed speech even with StarGAN. Speech intelligibility enhancement methods with only the Lombard effect still do not work well under strong noise interference with very low signal-to-noise ratios.

As shown in Figure 2, to address the challenges of complex communication scenarios, our previous work proposed a system called AdaSAStarGAN [24], which performs Lombard speech conversion based on different levels of near-end and far-end noise, for enhancing the intelligibility of near-end speech. The system consists of a generator based on the StarGAN framework and a discriminator, where the generator improves the intelligibility of the input speech and the discriminator evaluates the authenticity of the generated samples to guide the training of the generator. This approach is effective in optimizing intelligibility metrics. However, when we compare the waveform and spectrum of the source audio signal with those of the converted audio signal from the existing StarGAN-based model, we find that there is a sudden increase in amplitude and a shift of energy to the high-frequency region in the converted waveform. Since these high-amplitude frames represent important words, these frames should not be given excessive amplitude values in IENH, otherwise the speech quality will be severely degraded. This suggests that the

existing StarGAN framework is prone to artifacts; we believe this is because it does not focus on the transformation style as expected. This transformation style often leads to distortion of speech features, such as breakup or jitter, which compromises the overall audio quality.

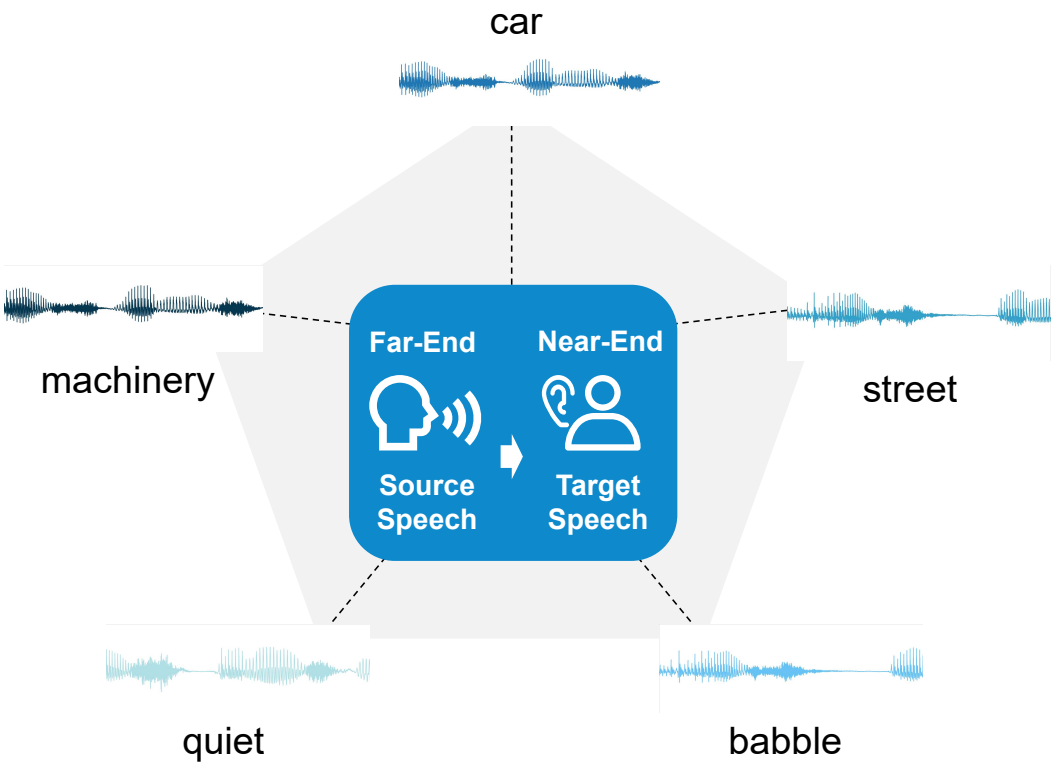

Figure 2. Diagram of Noise-adaptive feature conversion. The waveforms in the graph are clean speech with noise removed. The darker the waveform's color, the higher the decibel level of the noise. In addition, the lightest color is normal speech with 0dB of noise.

In this paper, we propose a StarGAN-based system called D2StarGAN as an extension of our earlier system to overcome the previously mentioned drawbacks. We employed a StarGAN-based optimization scheme that jointly maximizes multiple intelligibility metrics and quality metrics to improve speech quality and intelligibility. We performed a comprehensive evaluation of the performance of the system under different conditions. The experimental results show that the improved system significantly improves speech intelligibility and quality compared to the comparison method, outperforming the state-of-the-art baseline in objective and subjective assessments.

2. Related Work

According to different speech feature tuning strategies, IENH algorithms can be broadly classified into two main categories: rule-based IENH algorithms and data-driven IENH algorithms.

Rule-based algorithms [15,25–27] have been developed over decades in the field of speech processing and have accumulated a large number of models and techniques. These algorithms have the advantage of being fast to run, require no data training and are better able to account for variations in speech features. This makes them still advantageous in certain specific scenarios. However, the rule-based approach also has limitations in terms of speech intelligibility enhancement. Speech intelligibility enhancement involves a large number of interacting speech features, and it is difficult to fully capture and model the complex relationships between these features through a fixed set of rules. Therefore, a rule-based approach will always have its limitations. In addition, fixed-rule-based speech enhancement is often problematic in terms of speech quality and naturalness. As

the judgment of the model is based on predefined rules, the enhanced speech may lose some naturalness and not sound natural enough. In addition, speech quality may also deteriorate, producing noise, distortion or other undesirable effects. To overcome these problems, modern speech enhancement methods tend to use data-driven approaches such as deep learning to better model and improve the complexity of speech features.

The data-driven approach uses a large amount of speech data to build a feature mapping model from normal speech to Lombard speech by machine learning algorithms to achieve speech style conversion (SSC). In recent years, Bayesian–Gaussian mixture models (BGMM) [20,28], deep neural networks (DNNs) [29,30], recurrent neural networks (RNNs) and their variants such as long–short-term memory (LSTM) [31] networks have been widely used for mapping acoustic features. However, current parallel SSC methods require a parallel corpus of source and target speech and usually require a temporal alignment operation on the training data, an operation that may lead to some feature distortions. In addition, speakers usually speak slowly under the effect of the Lombard reflex, which further increases the difficulty of parallel SSC. Therefore, some more sophisticated algorithms are needed to implement the alignment operation when dealing with a non-parallel corpus to avoid the feature distortion problem. Therefore, researchers have started to explore non-parallel SSC approaches, i.e., parallel corpora that do not depend on source and target speech. In recent studies, the use of non-parallel speech style conversion methods has been encouraged in order to avoid temporal alignment operations. Some of these studies [21,22] combine variants of generative adversarial networks (GANs), such as cycle consistent generative adversarial networks (CycleGAN) [32–34] and StarGAN [35], to perform mapping of MFCC features without parallel speech and temporal alignment procedures. Compared with parallel speech style conversion methods, they have better intelligibility and naturalness and are able to learn many-to-many mappings. Among them, CycleGAN [36] can only learn one-to-one mappings, while StarGAN [37] can handle many-to-many mappings with different gender and attribute domains simultaneously. The use of CycleGAN allows for IENH augmentation, while the use of StarGAN allows for simultaneous consideration of the effect of gender differences on Lombard language features.

3. Baseline

3.1. Traditional NELE System Structure

Figure 3 shows the non-parallel SSC framework used by the latest method [32,33,35]. The system aims to convert the normal style input speech to the same Lombard style output speech. In the conversion process, a normal-to-Lombard speech conversion module is used. It comprises three key components: vocoder analysis, feature mapping and vocoder synthesis. First, the input speech signal is subjected to vocoder analysis to extract speech features. Then, the extracted speech features that are closely associated with the Lombard-style speech are transformed using a mapping model. Finally, the mapped features, along with the unmodified features, are then fed into the vocoder for synthesis. The vocoder takes these modified and unmodified features as input and combines them to generate the final output speech.

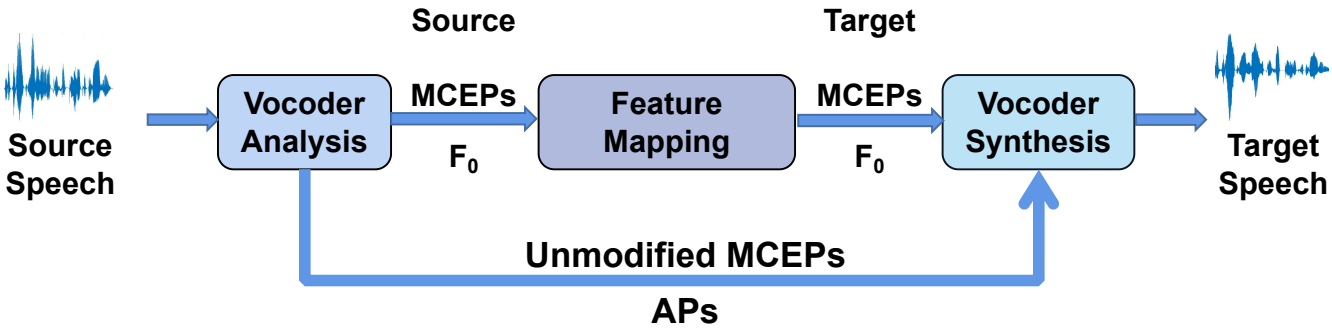

Figure 3. Schematic diagram of non-parallel SSC.

In this framework, there are various choices of vocoders. Commonly used parametric vocoders are STRAIGHT [38] and WORLD [39]. They extract three main features: (1) spectral envelope, which is usually expressed using mel-cepstral coefficients (MCEP); (2) fundamental frequency (F_0) and clear and turbid tone determination thresholds; and (3) non-periodic parameters (APs). In general, the non-periodic parameters are not modified, and the MCEPs as the main features can be partially or fully modified, depending on the system design.

The core module of the SSC framework is the mapping model, which needs to be pre-trained using the training data. In non-parallel SSC, the original data can be directly used for training without the need for temporal alignment operations for normal and Lombard speech with different durations. However, all baseline systems rely only on the Lombard effect, without differential mapping for proximal noise. These SOTA systems are not resistant to unstable interference from strong noise.

3.2. AdaSAStarGAN

Since the present method is an extension of our previously proposed AdaSAStarGAN, we first briefly review its formulation. The model diagram of AdaSAStarGAN is shown in Figure 4.

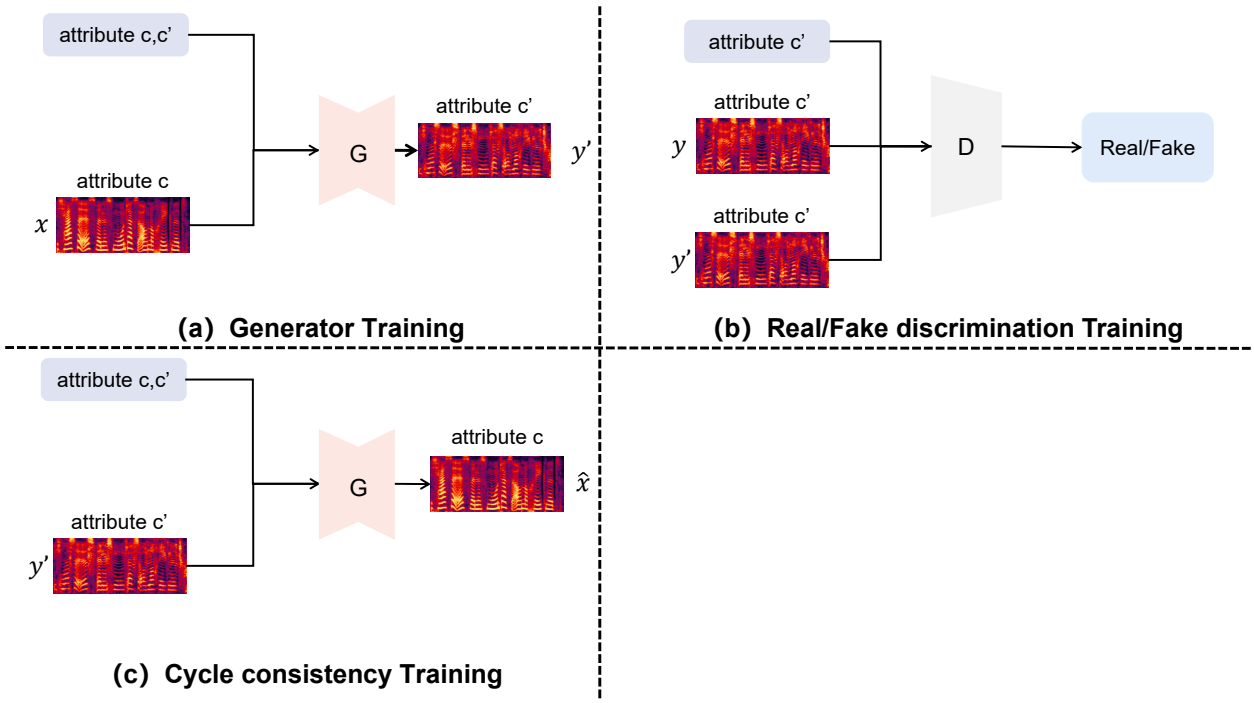

Figure 4. Feature Mapping of the AdaSAStarGAN: (**a**) **generator training:** receiving source speech x and label code c, c', generating enhanced speech $y' = G(x, c, c')$; (**b**) **real/fake discriminator training:** receiving enhanced speech y' or real speech y and label code c', discriminating the truthfulness of speech belonging to c'; (**c**) **cycle consistency training:** encouraging the D2StarGAN to retain in content-related features such that $G(y', c, c')$ is as close to x as possible.

The goal of training AdaSAStarGAN is to acquire a unified generator G that learns a mapping between multiple domains. More precisely, AdaSAStarGAN learns a generator G that converts the input acoustic features x into output features y', i.e., $G(x, c') \rightarrow y'$, conditional on the target domain label c′, and let y′ in $y' \in R^{Q \times T}$ be a sequence of acoustic features, where Q denotes the feature dimension and T represents the length of the sequence. The domain label code $c \in \{1, \ldots, C\}$ belongs to a set of possible labels, where C denotes the total number of domains. The purpose of AdaSAStarGAN is to make the generator G learn many-to-many mappings between different domains. Inspired by StarGAN,

AdaSAStarGAN converges the generator G and the discriminator D in a mutual adversarial process by combining adversarial loss, cycle consistency loss and identity mapping loss.

Adversarial loss is a loss function, also known as source–target conditional adversarial loss. It serves to make the transformed features generated by the generator indistinguishable from the true target features, thus improving the performance of the generator. In the adversarial loss, the discriminator is trained to differentiate the transformed features from the true target features, while the generator is trained to generate increasingly realistic transformed features in order to deceive the discriminator to the maximum extent possible. In this process, the discriminator and the generator confront each other until the features generated by the generator are indistinguishable from the true target features.

$$L_{ad} = E_{(x,c)\sim P(x,c),c'\sim P(c')}[logD(x,c',c)] + E_{(x,c)\sim P(x,c),c'\sim P(c')}[logD(G(x,c,c'),c,c')] \quad (1)$$

where $c'\sim P(c')$ denotes the randomly sampled labels from the real data label distribution P. D is a target-conditional discriminator. It attempts to learn the best decision boundary between the transformed features and the true target domain acoustic features given the source domain label code c and the target domain label code c' by maximizing the classification loss. In contrast, generator G tries to make the transformed features indistinguishable from the true target domain acoustic features labeled with c', thus minimizing the adversarial loss.

Domain classification loss is a distinction between the features generated by the generator G into a source feature domain and a target feature domain. In this case, generator G aims to generate features that are similar to the target domain in order to deceive discriminator D. By classifying the target features with the source features, discriminator D can better distinguish which features belong to the target domain. This approach helps to improve the performance of the generator G by allowing it to generate more realistic target features:

$$L_{cls} = E_{x\sim P(x),c'\sim P(c')}[logP_C(c'|G(x,c'))] \quad (2)$$

Circular consistency loss is proposed to address the problem where the transformed acoustic features may lose their input components. Although adversarial loss and classification loss can make the transformed acoustic features realistic and classifiable, they each prompt different transformations, resulting in a highly under-constrained mapping process. Therefore, cyclic consistency loss is introduced to ensure that the transformed features retain their input components:

$$L_{cyc} = E_{x,c,c'}[\left\|x - G(G(x,c'),c)\right\|_1] \quad (3)$$

The role of the cyclic consistency loss is to encourage the generative model G to learn an optimal mapping function such that the feature $G(G(x,c'),n) \to \hat{x}$ after two mappings can approximate the original feature x while retaining the input component, i.e., $\hat{x} \approx x$. In this process, the generative model considers both the input x, the source domain label code c, and the target domain label code c' while learning the mapping function to ensure that the generated features do not lose the information of the original input.

Identity mapping loss: To further restrict the generative model from changing features other than style during the transformation process, identity mapping loss is introduced. Specifically, the identity mapping loss encourages the generative model to learn a mapping function such that the input features can remain unchanged after mapping, i.e., $G(x,c) \approx x$. This ensures that the generative model does not lose important information other than style during the transformation process, thus improving the performance of the model:

$$L_{id} = E_{(x,c)\sim P(x,c)}[\|G(x,c) - x\|] \quad (4)$$

where D and G are optimized by minimizing adversarial loss, cyclic consistency loss, and identity mapping loss, respectively.

4. Proposed D2StarGAN

4.1. System Structure

As mentioned in Section 3.1 in previous work, we proposed a non-parallel SSC framework, AdaSAStarGAN, for adaptive speech enhancement in distal and proximal complex noisy environments. In practical scenarios, converting normal speech to Lombard speech faces limitations as it assumes a completely noise-free environment for the speaker to produce normal speech. However, in certain speech call situations where the speaker (far-end) and the receiver (near-end) are located in different noisy environments, the speaker may need to resort to Lombard speech for effective communication. In such cases, employing the conventional normal-to-Lombard conversion method can lead to a decline in the quality and comprehensibility of the converted speech. In addition, since more noise leads to a greater Lombard effect, i.e., there are different levels of Lombard speech, conversion between different levels of Lombard speech needs to be considered. To solve these problems, a framework is proposed that enables the conversion between normal speech to different levels of Lombard speech in different noise environments to improve speech quality and intelligibility.

To address the above limitations, we propose a speech intelligibility enhancement system, D2StarGAN, which adds a proximal noise processing module to the previous framework. This module extracts the features of the proximal noise signal and uses them as one of the inputs to the mapping model. In this way, we can make the generator not only focus on the features of the original speech but also flexibly adjust the enhanced speech to different noise types or the effect of non-smooth noise according to the features of the proximal noise. The framework diagram is shown in Figure 5. In the application scenario where both distal and proximal noises exist, we can use the distal noise level c' and the proximal noise level c as conversion conditions to achieve conversion between different levels of Lombard speech and normal speech, thus improving the intelligibility and quality of speech.

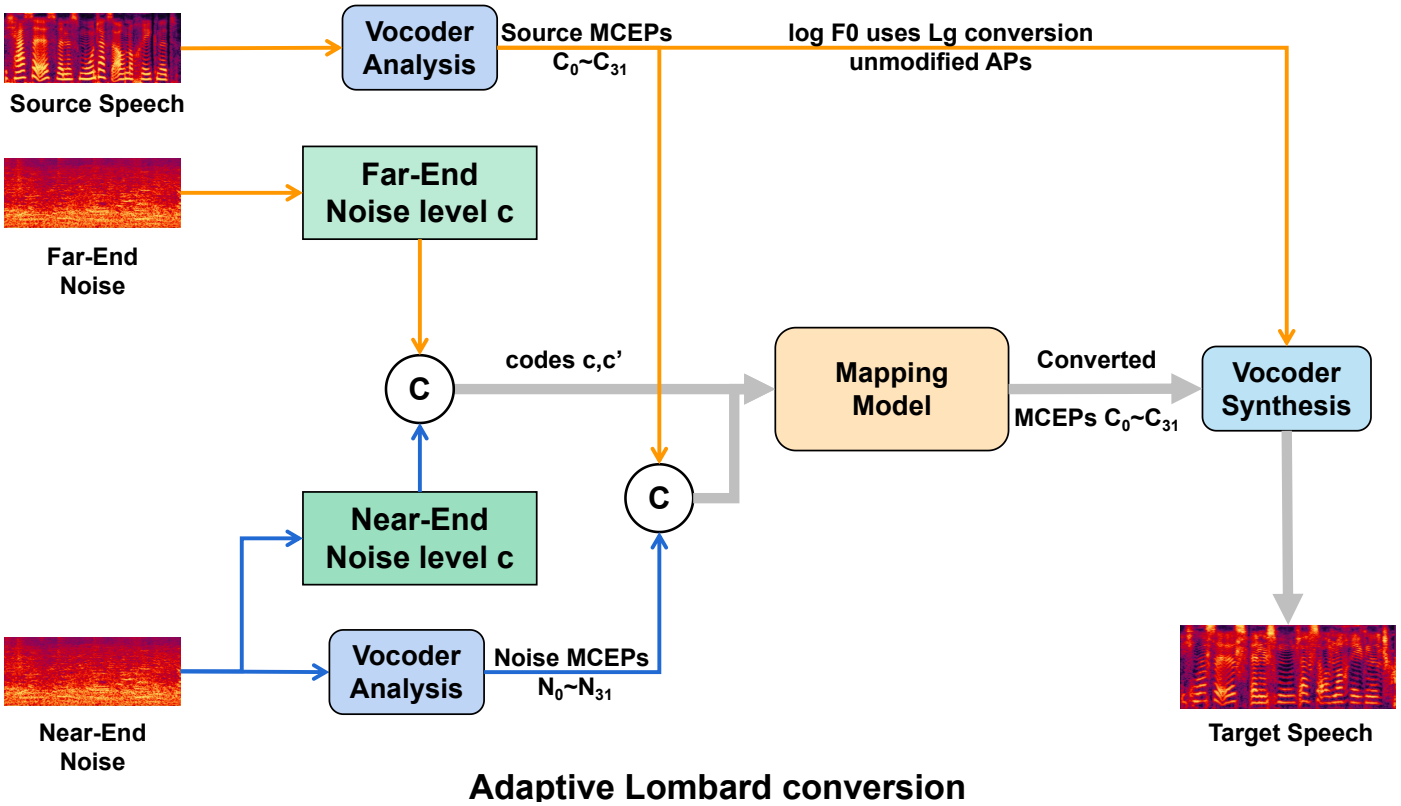

Figure 5. The overall framework of our proposed method.

The inputs to this framework include noisy mixed speech and the speaker's speech voice from the distal environment and noisy signals from the proximal environment. The

next processing steps include speech separation, sound pressure level extraction, speech feature extraction, preprocessing, feature mapping, and target speech synthesis. In the speech separation step, the distal speech is divided into distal noisy speech and clean speech as the source speech. In the sound pressure level extraction step, we measure the sound pressure levels of the proximal and distal noisy audio and label them with labels c and c'. In the speech feature extraction step, a WORLD [24] vocoder is used to extract features such as mel-inverse spectral coefficients (MCEPs), logarithmic fundamental frequencies (log F0), and acyclicity (APs) from the source speech and near-end noisy signals. the WORLD vocoder is analyzed and synthesized at a 5 ms frame shift to represent the spectral envelope as 32 dimensions ($C_0 \sim C_{31}$, $N_0 \sim N_{31}$) for feature mapping. Since F0 is a one-dimensional feature, it is difficult to train it together with other high-dimensional features (e.g., MCEP). Therefore, nonlinear transformation of F0 by logarithmic Gaussian normalization transform [35] is a widely used method. In the preprocessing stage, we normalize all features and then connect the distal and proximal noise levels, source speech features, and proximal noise features. In the feature mapping stage, the source speech features are mapped to the target enhanced speech features using the distal and proximal noise levels and the distal noise features as mapping conditions. In the target speech synthesis stage, the target-enhanced speech is synthesized using the WORLD vocoder. The final target enhanced speech is obtained.

4.2. Mapping Model

In this section, we present our proposed speech intelligibility enhancement system D2StarGAN, which adds a metric discriminator to previous work for optimizing speech metrics. Partly inspired by the NELE-GAN proposed by Haoyu Li [40], a time-frequency correlated amplification factor is obtained based on GAN for reassigning energy to the source speech.

Figure 6 shows a diagram of the D2StarGAN model of our proposed system. It consists of a generator G, a true–false discriminator ($D_{r/f}$) and an indicator discriminator (D_{ind}). g receives the input speech x and the proximal noise n and the distal and proximal noise level label codes c and c', where the domain label code of the input speech x is c. Then the output enhanced speech $y' = G(x, n, c')$. Next, Dr/f determines whether y' is true data belonging to the domain c', and Dind predicts the intelligibility and quality scores of the augmented speech. D_{ind}'s predicted scores are close to the true scores calculated for the target objective metric. Compared with those very complex original metrics, the gradients of the DNN-based discriminator can be easily computed and back-propagated to G. Therefore, under the guidance of Dr/f and D_{ind}, G can be efficiently trained to generate Lombard speech for the corresponding domain and optimize the learning metric of interest.

The training process can be divided into four modules: (a) Generator training, G takes three inputs: source speech x, proximal noise n and distal and proximal noise level labeling codes c and c'. Then G outputs the enhanced speech $y' = G(x, n, c')$. When training this part, if the target domain is normal speech in a noise-free environment, then the value of n is 0. (b) True–false discriminator training, D requires two inputs: the enhanced speech y' or y, and the label code c', and Dr/f outputs y' or y belonging to the c' domain truthfulness $D(y', c')$, which ranges from 0 to 1. (c) Cycle consistency training: Auxiliary D2StarGAN retains the content in x so that $G(y', n, c)$ is as close as possible to x. When training this part, we set the noise n as smooth noise because the input speech x is not noise adaptive. (d) Metric discriminator training. To predict intelligibility and quality scores, Dind requires three inputs: (1) augmented speech $G(x, n, c')$, (2) clean speech x, and (3) background noise n. We introduce the so-called I-function to represent the target metric to be modeled (described later). In addition, other reference algorithms (e.g., NELE-GAN) pre-enhanced signal examples y' are also fed into the training D_{ind} to stabilize the training process and improve the results. In the inference stage, we use the trained generator G for feature mapping.

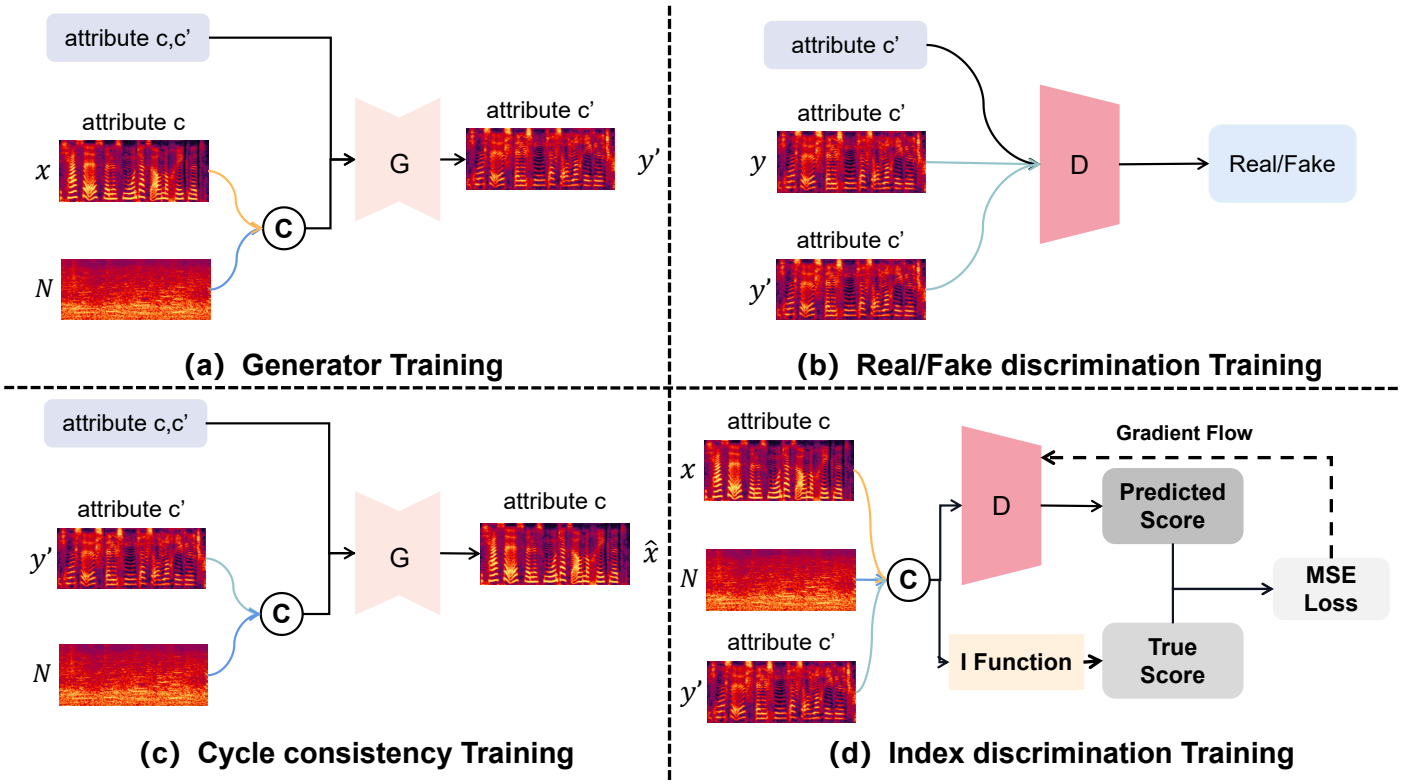

Figure 6. Diagram of the D2StarGAN: (**a**) **generator training:** receiving source speech x, noise n and label code c, c', generating enhanced speech $y' = G(x, n, c')$; (**b**) **real/fake discriminator training:** receiving enhanced speech y' or real speech y and label code c', discriminating the truthfulness of speech belonging to c'; (**c**) **cycle consistency training:** encouraging the D2StarGAN to retain content-related features such that $G(y', n, c)$ is as close to x as possible; (**d**) **index discriminator training:** predicting intelligibility and quality scores, receiving enhanced speech $G(x, n, c')$, clean speech x and background noise n, and introducing I function to represent the target metric model.

Adversarial loss: Adversarial loss function in D2StarGAN is designed to make the transformed features generated by the generator indistinguishable from the true target features in the discriminator D. The expression for this loss function is as follows:

$$L_{ad} = E_{x,c'}[logD_{r/f}(x, c')] + E_{x,n,c'}[log(1 - D_{r/f}(G(x, n, c'), c')] \tag{5}$$

is the adversarial loss of the generator G. $\mathcal{L}_{ad}$ takes a larger value when D correctly classifies $G(x, c)$ and y as false and true speech features, while $\mathcal{L}_{ad}$ takes a smaller value when G successfully deceives D. Thus, $G(x, c)$ is misclassified by D as a true speech feature. Therefore, we want to maximize $\mathcal{L}_{ad}$ with respect to D and minimize Lad with respect to G.

The loss function of D_{ind} is represented as follows:

$$\begin{aligned}\mathcal{L}_D^{ind} = E_{x,n,c,c'}\{&[D_{ind}(G(\boldsymbol{x}, n, c'), \boldsymbol{x}, n) - I(G(\boldsymbol{x}, n, c'), \boldsymbol{x}, n)]^2 \\ &+ [D_{ind}(\widehat{y}, \boldsymbol{x}, n) - I(\widehat{y}, \boldsymbol{x}, n)]^2\}\end{aligned} \tag{6}$$

By minimizing $\mathcal{L}_D^{ind}$, D_{ind} is encouraged to accurately predict speech intelligibility and quality scores.

We fix the parameters of the D_{ind} first, then we apply a back-propagated gradient to update G to maximize the predicted scores. In order to maximize the predicted scores, we use the following loss function:

$$\mathcal{L}_G^{ind} = E_{x,n,c'}[D_{ind}(G(\boldsymbol{x}, n, c'), \boldsymbol{x}, n) - t]^2 \tag{7}$$

where t denote the maximum scores of the intelligibility and quality metrics.

Domain classification loss is used to generate features in the target feature domain for the optimal discriminator D:

$$\mathcal{L}_{cls} = E_{x,n,c'}[logP_C(c'|G(x,n,c'))] \tag{8}$$

Cycle consistency loss is used to ensure that the transformed features will maintain the composition of the input, as shown by the following equation:

$$\mathcal{L}_{cyc} = E_{x,c,c'}[\left\|x - G(G(x,n,c'),\overline{n},c)\right\|_1] \tag{9}$$

The goal of generator G is to ensure, as far as possible, that the source data after conversion to c' still has the original features when mapped back to the original domain c. The L_1 parametrization here represents the difference between the source data before and after the transformation, and the optimization goal of generator G is to minimize this difference.

Optimization by minimizing $L_{ad}, L_G^{ind}, L_{cyc}, L_{id}$ and L_{cls}. Discriminator D optimization by minimizing $L_{ad}, L_G^{ind}, L_{cyc}, L_{id}$ and L_{cls}. The parameters of the total loss function for losses other than the adversarial loss are set to $\lambda_{cyc} = 10$, $\lambda_{id} = 5$,$\lambda_{cls} = 1$.

I Function: I Function in the framework is mainly composed of the following three metrics, all of which are the most widely used speech intelligibility metrics: (1) SIIB: The Speech Intelligibility in Bits metric calculates the amount of information shared between clean and distorted speech signals, measured in bits per second. (2) ESTOI: Extended short-time objective intelligibility measures intelligibility by computing the correlation between the spectra of clean and distorted speech. (3) HASPI: Hearing-aid speech perception index evaluates intelligibility loss by examining cepstral correlation and auditory coherence within an auditory model.

Network Architectures: (1) *Generator and Real/Fake Discriminator*: The details of the Generator architectures are given in Figure 7a. We designed the network architectures based on our previous work. The input features for G are extracted from input speech and near-end noise. We used a 2-1-2D CNN in G and a 2D CNN in $D_{r/f}$. The self-attention layer [41] is used in the 1D residual block of G to improve the fitting ability of the model. For a GAN objective, we used the least squares GAN. IN, AdaIN, GLU and PS indicate instance normalization, adaptive instance normalization, gated linear unit, and pixel shufflert, respectively. The generator is *fully convolutional* (FC). We trained the networks using the Adam optimizer with a batch size of 1, in which we used a randomly cropped segment (128 frames) as one instance. During training, we normalize all metric scores to the range [0, 1], the same range as the sigmoid activation, and set the target maximum score to the maximum score plus 0.1 obtained by the D_{ind} at the epoch. The number of iterations was set to 1×10^5, the learning rates of G, D, and D_{ind} were set to 0.0002, 0.0001, and 0.0002, respectively. G was trained once in 5 iterations and D_{ind} was trained once in 100 iterations. We set $\lambda_{cyc} = 10$, $\lambda_{id} = 5$, $\lambda_{ind} = 2.5$.

(2) *Index Discriminator*: The details of the Index Discriminator architectures are given in Figure 7b. The input to the index discriminator is the concatenated MFCCs of enhanced speech, unenhanced speech, and noisy audio. Each convolution layer has an LReLU activation. A global average pooling (GAPool) is added after the last CNN module. The last FC layer with sigmoid activation predicts the scores of all metrics.

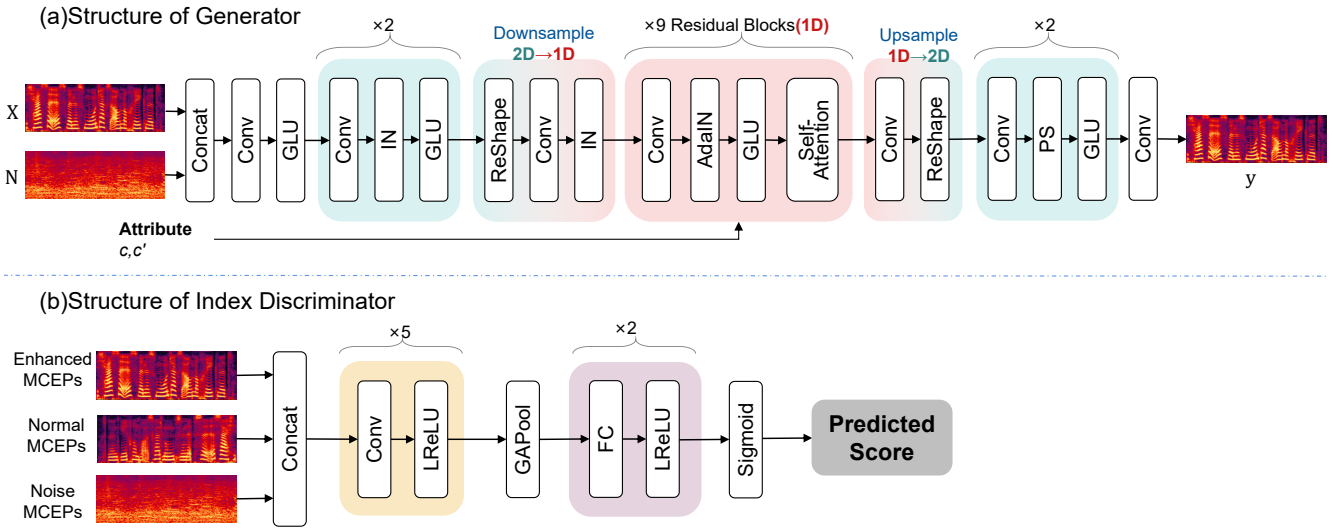

Figure 7. Network architectures of our D2StarGAN model. (**a**) Structure of the generator; (**b**) structure of the index discriminator.

5. Experimental Section and Results

5.1. Experimental Setup

5.1.1. Dataset

Lombard speech database for German language [42] is used for evaluating the method on the IENH task. This database consists of recordings of both normal speech in a noiseless environment and Lombard speech in low- and high-noise environments. There are separate subsets for training and evaluation. The training subset contains 256 sentences, while the evaluation subset consists of 64 sentences. Both subsets include recordings of normal speech in a noiseless environment as well as two levels of Lombard speech recorded in low- and high-noise environments. To standardize the data, the recordings were downsampled to a sampling rate of 16 kHz for the purpose of this study.

NOISEX-92: [43] We used five background noises from the NOISEX-92 dataset, including station, canteen, white, babble and restaurant noise. Of these, station, canteen and white noise were chosen for the training and validation data, while white noise was only used in the loop consistent training. The remaining two noises were used for the test data. We produce low-SNR versions of babble and restaurant noise and high-SNR versions as low noise and high noise.

Specifically, the training set contained a corpus of 1536 sentences (256 sentences × 3 levels × 2 noises), while the test set contained a corpus of 384 sentences (64 sentences × 3 levels × 2 noises). The corpus in the test set covered 10 different auditory conditions, including two ends, three levels and two noise types. In the near-end noise-free condition, the noise type was not considered. We use a mixture of recordings from the noise dataset as well as the speech dataset to simulate the far and near end of the conversation

5.1.2. Comparison with Existing Approaches

Three latest non-parallel SSC methods were involved in the experiments. They included a CycleGAN-based method (CGAN) [4], a StarGAN-based method (SGAN) [15] and an AdaSAStarGAN-based method (ASSGAN) [32]. UN indicates speech without any modification.

5.2. Objective Evaluations

In our study, we used five target metrics (SIIB, HASPI, ESTOI, SRMR, NCM) to assess the intelligibility of the system. Notably, two advanced metrics, SRMR [44] and NCM [45], were incorporated into the evaluation, which was not utilized during the model's pre-training phase. Speech-to-reverberation modulation energy ratio (SRMR) is a

metric that evaluates speech quality and intelligibility based on the modulation spectral representation of the speech signal, while normalized-covariance measure (NCM) is based on the modulation spectral representation of the speech signal and the corresponding reverberant signal. Higher scores in these metrics indicate better system performance in terms of intelligibility. The three metrics (SIIB, HASPI, and ESTOI) involved in pre-training were complemented by the two new metrics (SRMR and NCM) during evaluation.

In Tables 1 and 2, we present the average objective scores for each system in Babble and Restaurant noise conditions. Furthermore, Table 3 ranks the scores of each system at various noise levels, ranging from lowest to highest. The results from these tables highlight that the D2StarGAN system clearly outperforms the state-of-the-art baseline system in terms of intelligibility, achieving higher scores. This improvement can be attributed to the incorporation of new target metrics and the utilization of an advanced network structure. Specifically, D2StarGAN demonstrated superior performance on previously unseen SRMR and NCM scores, providing further evidence that employing multi-metric optimization strategies can effectively enhance intelligibility.

Table 1. Average objective scores of the compared systems under different noise types under noise **Babble** noise. Noiseless, low noise and high noise represent environmental noise levels. The noise level in front of the symbol "→" indicates the far-end noise level, and the noise level after it indicates the near-end noise level.

System	Noiseless → Low Noise					Noiseless → High Noise				
	SIIB	**HASPI**	**ESTOI**	**SRMR**	**NCM**	**SIIB**	**HASPI**	**ESTOI**	**SRMR**	**NCM**
UN	18.71	1.92	0.24	3.76	0.31	13.08	1.58	0.20	2.17	0.20
CGAN	36.21	2.58	0.32	4.13	0.40	23.86	2.08	0.24	4.39	0.33
SGAN	40.49	2.64	0.37	6.42	0.44	**32.17**	2.29	0.25	5.42	0.34
ASSGAN	47.25	2.66	0.37	6.63	0.43	30.38	2.25	0.25	5.33	0.33
D2StarGAN	**49.56**	**2.70**	**0.39**	**7.43**	**0.48**	31.57	**2.31**	**0.27**	**6.16**	**0.37**

Table 2. Average objective scores of the compared systems under different near-end noise conditions under **Restaurant** noise.

System	Noiseless → Low Noise					Noiseless → High Noise				
	SIIB	**HASPI**	**ESTOI**	**SRMR**	**NCM**	**SIIB**	**HASPI**	**ESTOI**	**SRMR**	**NCM**
UN	15.86	1.66	0.21	2.23	0.28	10.94	1.49	0.13	1.42	0.19
CGAN	25.12	2.13	0.28	4.59	0.36	16.84	1.83	0.19	2.68	0.28
SGAN	31.58	2.47	0.30	5.20	0.44	20.83	2.00	0.19	3.32	0.32
ASSGAN	36.82	2.34	0.29	5.39	0.45	21.28	1.97	0.20	3.37	0.33
D2StarGAN	**39.90**	**2.53**	**0.31**	**5.83**	**0.47**	**23.13**	**2.01**	**0.21**	**3.77**	**0.36**

Table 3. Average objective scores of the compared systems under different noise types under **noise Low noise → High noise** condition. Tests were conducted using Babble and Restaurant with low as well as high SNR, respectively.

System	Babble					Restaurant				
	SIIB	**HASPI**	**ESTOI**	**SRMR**	**NCM**	**SIIB**	**HASPI**	**ESTOI**	**SRMR**	**NCM**
UN	14.31	1.61	0.21	2.23	0.24	9.37	1.43	0.14	1.43	0.19
SGAN	33.30	2.37	0.26	5.33	0.37	20.66	**2.05**	0.20	3.30	0.35
ASSGAN	32.59	2.38	0.26	5.59	0.37	21.37	2.03	0.20	3.31	0.35
D2StarGAN	**33.31**	**2.42**	**0.28**	**6.14**	**0.39**	**22.37**	**2.05**	**0.21**	**3.64**	**0.37**

5.3. Subjective Listening Tests

In the conducted subjective listening evaluation, participants used the Comparative Mean Opinion Score (CMOS) standard to compare the speech quality of different ap-

proaches. They rated the approaches on a scale from −3 to 3, ranging from "much worse" to "much better", with 0 indicating no significant difference ("about the same").

The evaluation focused on two key aspects: intelligibility and naturalness. To ensure reliable results, utterances that did not achieve an average word accuracy of at least 60% were excluded from the assessment. A total of 20 participants, aged between 20 and 40, took part in the evaluation, each listening and evaluating 64 recordings, which consisted of four methods per utterance with two female and two male voices, leading to four comparison groups. The evaluation was conducted in an anechoic chamber using high-quality Audio-Technica ATH-M50x headphones. Speech signals were processed and mixed with noise to simulate various scenarios. A total of four different noise scenarios were included in the evaluation, and separate tests were performed for normal speech, low Lombard speech, and high Lombard speech. The average values of these tests were used in the final results. The comparison method involved using the model of normal speech converted to high Lombard speech.

The results presented in Figure 8 show the average CMOS scores of the D2StarGAN approach compared to the baselines. Notably, all scores were above 0, indicating that the D2StarGAN consistently outperformed the baselines. When compared to unprocessed speech, the D2StarGAN approach received scores around 2, indicating a significant improvement in speech intelligibility and naturalness (IENH). Moreover, when compared to the CGAN, SGAN, and ASSGAN approaches, the D2StarGAN scores ranged from 0.6 to 0.9, corresponding to "slightly better" ratings.

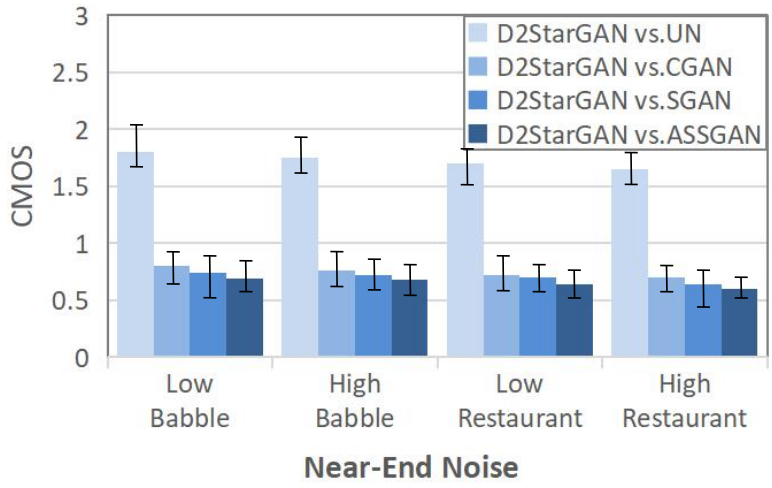

Figure 8. CMOS results with 95% confidence intervals.

5.4. Acoustic Analysis

The acoustic properties of the enhanced speech were analyzed in detail and studied in comparison with AdaSAStarGAN. In Figure 9, examples of the waveforms and spectrograms of the different signals are shown. By looking at the spectrograms, we can see that both AdaSAStarGAN and D2StarGAN modify the characteristics of the speech signal by redistributing the energy. Comparing Figure 9a,b, we can see that D2StarGAN tends to allocate more energy in the mid-frequency region of the voiced segment (inside the orange dashed box), while AdaSAStarGAN focuses more on the high-frequency region of the unvoiced segment (inside the blue dashed box). Furthermore, we observed that the D2StarGAN-enhanced speech waveform envelope was similar to the original unmodified speech, while the AdaSAStarGAN-modified waveform changed significantly, resulting in more acoustic artifacts.

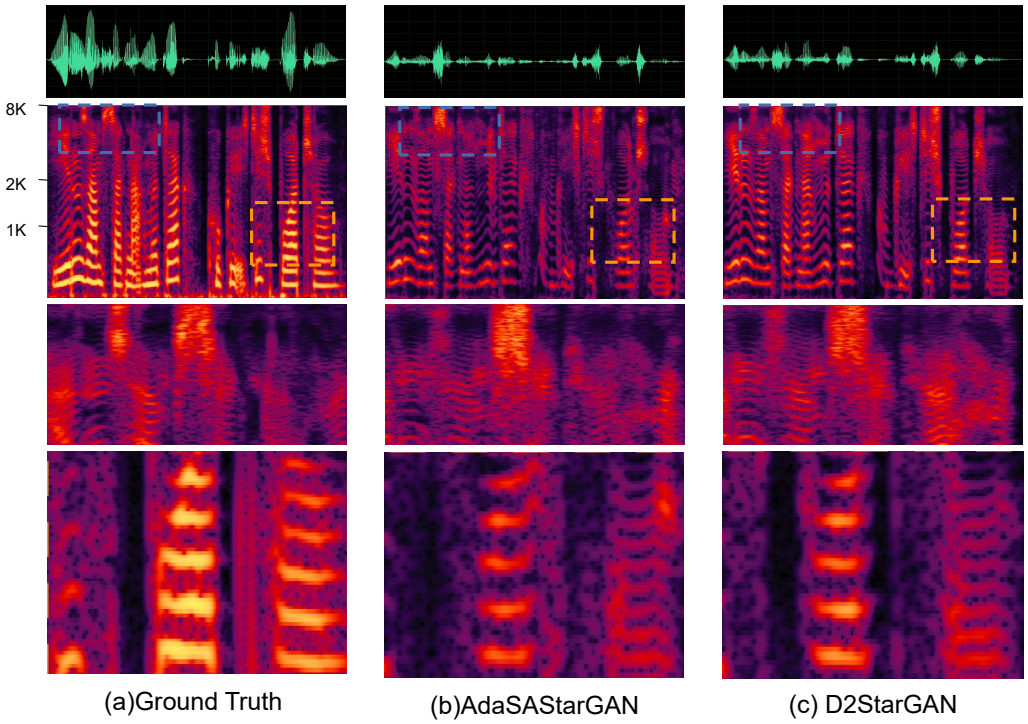

Figure 9. Waveforms and their spectrograms on one utterance in level 55 of Lombard: (**a**) Ground Truth, (**b**) enhanced speech from AdaSAStarGAN, and (**c**) enhanced speech from Proposed (All). The utterance used is notated as "f1_s34".

5.5. Analysis of System Robustness

We further analyzed the robustness of the system in the absence of visible speakers and languages. We used the Lombard Grid dataset [46] of English speakers to test our proposed system to verify its effectiveness under the mismatched speaker and language conditions. The Lombard Grid is an audiovisual Lombard speech corpus. The corpus comprises 54 speakers with 100 discourses each, including 50 Lombard and 50 normal discourses. In this experiment, we used our model trained on the German Lombard speech dataset and evaluated it on the Lombard Grid dataset with the same noise conditions as the original test set.

In this study, we conducted tests on a dataset that differs from the training set in terms of both language and speakers. As shown in Table 4, comparing our method to previous approaches, we consistently achieved favorable results, demonstrating the effectiveness of our approach in cross-lingual and cross-speaker scenarios.

Despite the differences in language and speaker characteristics between the training and test datasets, the performance of our system remains stable. This demonstrates the adaptability and versatility of our approach as it is able to handle a wide range of languages and speakers efficiently.

Table 4. Average Objective Scores On Lombard Grid Test Set.

System	SIIB	HASPI	ESTOI	SRMR	NCM
UN	13.64	1.60	0.21	2.22	0.23
SGAN	27.12	2.24	0.23	5.01	0.32
D2StarGAN	**29.85**	**2.29**	**0.26**	**5.83**	**0.35**

6. Conclusions

In this paper, aiming at solving the problem where the relatively smooth Lombard speech generated by existing speech intelligibility transformation frameworks is insufficient to meet real-life complex communication requirements, a metric-optimized near-end noise

adaptive speech intelligibility enhancement method is proposed, called D2StarGAN. This approach combines the concepts of StarGAN and a dual non-parallel SSC discriminator with a data-based strategy while taking into account the differences in complex telephonic speech environments. A dual non-parallel SSC discriminator is introduced to optimize the solvability metric in the D2StarGAN architecture. This discriminator can accurately distinguish the difference features between the converted speech and the target speech, resulting in more accurate conversion. We conducted a comprehensive objective and subjective evaluation of the proposed D2StarGAN approach. For the objective evaluation, we used a series of objective metrics to measure the quality of speech conversion, including speech intelligibility, sound quality retention, and speech similarity. For the subjective evaluation, we conducted a manual auditory assessment, inviting reviewers to rate the perception of the converted speech and provide feedback on their subjective opinions. The results show that our D2StarGAN method exhibits superior performance in both objective and subjective evaluations. Compared to traditional methods, D2StarGAN can achieve differential conversion more accurately in complex telephone speech environments, improving the interpretability and quality of speech. However, the current speech intelligibility enhancement technology remains far from the high comfort levels of practical applications. Future work will see further research conducted on the comfort and stability of practical applications.

Author Contributions: Conceptualization, D.L., C.Z. and L.Z.; Methodology, D.L.; Software, L.Z.; Writing—original draft, C.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Natural Science Foundation of China (U22A2035, No. 61701194), Application Foundation Frontier Special Project of Wuhan Science and Technology Plan Project (No. 2020010601012288), Doctoral Research Foundation of Jianghan University (2019029), Nature Science Foundation of Hubei Province (2017CFB756).

Data Availability Statement: The data that support the findings of this study are available in "Lombard speech database for german language" at http://doi.org/10.5281/zenodo.48713. The data that support the findings of this study are available in "Lombard Grid" at http://doi.org/10.1121/1.5042758, reference number EL523–EL529.

Conflicts of Interest: The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Li, G.; Hu, R.; Wang, X.; Zhang, R. A near-end listening enhancement system by RNN-based noise cancellation and speech modification. *Multimed. Tools Appl.* **2019**, *78*, 15483–15505. [CrossRef]
2. Leglaive, S.; Alameda-Pineda, X.; Girin, L.; Horaud, R. A recurrent variational autoencoder for speech enhancement. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 371–375.
3. Yemini, Y.; Chazan, S.E.; Goldberger, J.; Gannot, S. A Composite DNN Architecture for Speech Enhancement. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 841–845.
4. Kleijn, W.B.; Crespo, J.B.; Hendriks, R.C.; Petkov, P.; Sauert, B.; Vary, P. Optimizing speech intelligibility in a noisy environment: A unified view. *IEEE Signal Process. Mag.* **2015**, *32*, 43–54. [CrossRef]
5. Hussain, A.; Chetouani, M.; Squartini, S.; Bastari, A.; Piazza, F. Nonlinear speech enhancement: An overview. In *Progress in Nonlinear Speech Processing*; Springer: Berlin/Heidelberg, Germany, 2007; pp. 217–248.
6. Huang, P.S.; Chen, S.D.; Smaragdis, P.; Hasegawa-Johnson, M. Singing-voice separation from monaural recordings using robust principal component analysis. In Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, 25–30 March 2012; pp. 57–60.
7. Ephraim, Y.; Malah, D. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Trans. Acoust. Speech Signal Process.* **1985**, *33*, 443–445. [CrossRef]
8. Kwan, C.; Chu, S.; Yin, J.; Liu, X.; Kruger, M.; Sityar, I. Enhanced speech in noisy multiple speaker environment. In Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, China, 1–8 June 2008; pp. 1640–1643.
9. Lu, X.; Tsao, Y.; Matsuda, S.; Hori, C. Speech enhancement based on deep denoising autoencoder. In Proceedings of the Interspeech, Lyon, France, 25–29 August 2013; Volume 2013, pp. 436–440.

10. Kolbæk, M.; Tan, Z.H.; Jensen, J. Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2016**, *25*, 153–167. [CrossRef]
11. Fu, S.W.; Wang, T.W.; Tsao, Y.; Lu, X.; Kawai, H. End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2018**, *26*, 1570–1584. [CrossRef]
12. Sun, L.; Du, J.; Dai, L.R.; Lee, C.H. Multiple-target deep learning for LSTM-RNN based speech enhancement. In Proceedings of the 2017 Hands-Free Speech Communications and Microphone Arrays (HSCMA), San Francisco, CA, USA, 1–3 March 2017; pp. 136–140.
13. Ayhan, B.; Kwan, C. Robust speaker identification algorithms and results in noisy environments. In Proceedings of the Advances in Neural Networks–ISNN 2018: 15th International Symposium on Neural Networks, ISNN 2018, Minsk, Belarus, 25–28 June 2018; Proceedings 15; Springer: Berlin/Heidelberg, Germany, 2018; pp. 443–450.
14. Huang, Z.; Watanabe, S.; Yang, S.w.; García, P.; Khudanpur, S. Investigating self-supervised learning for speech enhancement and separation. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; pp. 6837–6841.
15. Zorila, T.C.; Kandia, V.; Stylianou, Y. Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression. In Proceedings of the Thirteenth Annual Conference of the International Speech Communication Association, Portland, OR, USA, 9–13 September 2012.
16. Jokinen, E.; Remes, U.; Takanen, M.; Palomäki, K.; Kurimo, M.; Alku, P. Spectral tilt modelling with extrapolated GMMs for intelligibility enhancement of narrowband telephone speech. In Proceedings of the 2014 14th International Workshop on Acoustic Signal Enhancement (IWAENC), Juan-les-Pins, France, 8–11 September 2014; pp. 164–168.
17. Garnier, M.; Henrich, N. Speaking in noise: How does the Lombard effect improve acoustic contrasts between speech and ambient noise? *Comput. Speech Lang.* **2014**, *28*, 580–597. [CrossRef]
18. Junqua, J.C.; Fincke, S.; Field, K. The Lombard effect: A reflex to better communicate with others in noise. In Proceedings of the 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258), Phoenix, AZ, USA, 15–19 March 1999; Volume 4, pp. 2083–2086.
19. Jokinen, E.; Remes, U.; Alku, P. Intelligibility enhancement of telephone speech using Gaussian process regression for normal-to-Lombard spectral tilt conversion. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *25*, 1985–1996. [CrossRef]
20. Li, G.; Wang, X.; Hu, R.; Zhang, H.; Ke, S. Normal-to-lombard speech conversion by LSTM network and BGMM for intelligibility enhancement of telephone speech. In Proceedings of the 2020 IEEE International Conference on Multimedia and Expo (ICME), London, UK, 6–10 July 2020; pp. 1–6.
21. Kaneko, T.; Kameoka, H.; Tanaka, K.; Hojo, N. Stargan-vc2: Rethinking conditional methods for stargan-based voice conversion. *arXiv* **2019**, arXiv:1907.12279.
22. Ferro, R.; Obin, N.; Roebel, A. Cyclegan voice conversion of spectral envelopes using adversarial weights. In Proceedings of the 2020 28th European Signal Processing Conference (EUSIPCO), Amsterdam, The Netherlands, 18–22 January 2021; pp. 406–410.
23. Li, H.; Fu, S.W.; Tsao, Y.; Yamagishi, J. iMetricGAN: Intelligibility enhancement for speech-in-noise using generative adversarial network-based metric learning. *arXiv* **2020**, arXiv:2004.00932.
24. Li, D.; Zhao, L.; Xiao, J.; Liu, J.; Guan, D.; Wang, Q. Adaptive Speech Intelligibility Enhancement for Far-and-Near-end Noise Environments Based on Self-attention StarGAN. In *International Conference on Multimedia Modeling*; Springer: Cham, Switzerland, 2022; pp. 205–217.
25. Sauert, B.; Vary, P. Near end listening enhancement: Speech intelligibility improvement in noisy environments. In Proceedings of the 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, Toulouse, France, 14–19 May 2006; Volume 1, p. I.
26. Koutsogiannaki, M.; Petkov, P.N.; Stylianou, Y. Intelligibility enhancement of casual speech for reverberant environments inspired by clear speech properties. In Proceedings of the Sixteenth Annual Conference of the International Speech Communication Association, Dresden, Germany, 6–10 September 2015.
27. Niermann, M.; Vary, P. Listening Enhancement in Noisy Environments: Solutions in Time and Frequency Domain. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *29*, 699–709. [CrossRef]
28. López, A.R.; Seshadri, S.; Juvela, L.; Räsänen, O.; Alku, P. Speaking Style Conversion from Normal to Lombard Speech Using a Glottal Vocoder and Bayesian GMMs. In Proceedings of the Interspeech, Stockholm, Sweden, 20–24 August 2017; pp. 1363–1367.
29. Seshadri, S.; Juvela, L.; Räsänen, O.; Alku, P. Vocal effort based speaking style conversion using vocoder features and parallel learning. *IEEE Access* **2019**, *7*, 17230–17246. [CrossRef]
30. Li, G.; Hu, R.; Zhang, R.; Wang, X. A mapping model of spectral tilt in normal-to-Lombard speech conversion for intelligibility enhancement. *Multimed. Tools Appl.* **2020**, *79*, 19471–19491. [CrossRef]
31. Gentet, E.; David, B.; Denjean, S.; Richard, G.; Roussarie, V. Neutral to lombard speech conversion with deep learning. In Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 7739–7743.
32. Seshadri, S.; Juvela, L.; Yamagishi, J.; Räsänen, O.; Alku, P. Cycle-consistent adversarial networks for non-parallel vocal effort based speaking style conversion. In Proceedings of the ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 6835–6839.

33. Seshadri, S.; Juvela, L.; Alku, P.; Räsänen, O. Augmented CycleGANs for Continuous Scale Normal-to-Lombard Speaking Style Conversion. In Proceedings of the Interspeech, Graz, Austria, 15–19 September 2019; pp. 2838–2842.
34. Xiao, J.; Liu, J.; Li, D.; Zhao, L.; Wang, Q. Speech Intelligibility Enhancement By Non-Parallel Speech Style Conversion Using CWT and iMetricGAN Based CycleGAN. In Proceedings of the MultiMedia Modeling: 28th International Conference, MMM 2022, Phu Quoc, Vietnam, 6–10 June 2022; Proceedings, Part I; Springer: Cham, Switzerland, 2022; pp. 544–556.
35. Li, G.; Hu, R.; Ke, S.; Zhang, R.; Wang, X.; Gao, L. Speech intelligibility enhancement using non-parallel speaking style conversion with stargan and dynamic range compression. In Proceedings of the 2020 IEEE International Conference on Multimedia and Expo (ICME), London, UK, 6–10 July 2020; pp. 1–6.
36. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
37. Choi, Y.; Choi, M.; Kim, M.; Ha, J.W.; Kim, S.; Choo, J. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8789–8797.
38. Kawahara, H.; Masuda-Katsuse, I.; De Cheveigne, A. Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. *Speech Commun.* **1999**, *27*, 187–207. [CrossRef]
39. Morise, M.; Yokomori, F.; Ozawa, K. WORLD: A vocoder-based high-quality speech synthesis system for real-time applications. *IEICE Trans. Inf. Syst.* **2016**, *99*, 1877–1884. [CrossRef]
40. Li, H.; Yamagishi, J. Multi-Metric Optimization Using Generative Adversarial Networks for Near-End Speech Intelligibility Enhancement. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 3000–3011. [CrossRef]
41. Phan, H.; Le Nguyen, H.; Chén, O.Y.; Koch, P.; Duong, N.Q.; McLoughlin, I.; Mertins, A. Self-attention generative adversarial network for speech enhancement. In Proceedings of the ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual, 6–12 June 2021; pp. 7103–7107.
42. Soloducha, M.; Raake, A.; Kettler, F.; Voigt, P. Lombard speech database for German language. In Proceedings of the DAGA 42nd Annual Conference on Acoustics, Florence, Italy, 24–27 October 2016.
43. Varga, A.; Steeneken, H.J. Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Commun.* **1993**, *12*, 247–251. [CrossRef]
44. Falk, T.H.; Zheng, C.; Chan, W.Y. A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. *IEEE Trans. Audio Speech Lang. Process.* **2010**, *18*, 1766–1774. [CrossRef]
45. Ma, J.; Hu, Y.; Loizou, P.C. Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. *J. Acoust. Soc. Am.* **2009**, *125*, 3387–3405. [CrossRef] [PubMed]
46. Alghamdi, N.; Maddock, S.; Marxer, R.; Barker, J.; Brown, G.J. A corpus of audio-visual Lombard speech with frontal and profile views. *J. Acoust. Soc. Am.* **2018**, *143*, EL523–EL529. [CrossRef] [PubMed]