1. INTRODUCTION

In a typical noisy environment, an audio signal is perceived as a mixture between a desired audio (speech) and an intrusive background noise. Accordingly, speech enhancement or denoising is interpreted as a source separation problem, where the goal is to separate the desired audio signal from the intrusive noise. The background noise type and the signal-to-noise ratio (SNR) have a direct influence on the quality of the estimated denoised speech. For instance, some common background noise types can be very similar to the desired speech such as cafe or food court noise. In these cases, estimating the desired speech from the mixed is challenging and sometimes impossible in low SNR situations as the components of the mixed signal are sharing the same distribution and frequency bands. This process of eliminating background noise from speech tracks is a vital building block for automatic speech recognition (ASR) systems.

Previously, traditional approaches were adopted for speech enhancement applications such as spectral subtraction [12] and binary masking techniques [3, 4]. Moreover, statistical approaches based on variants of Wiener filter were also applied in speech enhancement [5, 6, 7]. However, all of these approaches require a prior knowledge of the SNR and can operate on limited noise types in high SNR situations.

To overcome such SNR assumptions and noise type limitations, data driven approaches based on deep neural networks (DNNs) are widely used in literature to learn deep underlying features of either the required speech or the intrusive background noise from the given data. For instance, denoising autoencoders (AE) were used in [8, 9] to estimate a clean track from a noisy input based on the L1-loss. Long short-term memory (LSTM) networks have also been utilized to incorporate temporal speech structure in the denoising process [10, 11]. Also, an adaptation of the autoregressive generative WavNet was used in [12] where a denoised sample is generated based on the previous input and output denoised samples.

In 2014, generative adversarial networks (GANs) were introduced as the state-of-the-art for deep generative models [13]. In GANs, a generator is trained adversarially with a discriminator to generate images belonging to the same joint distribution of the training data. Afterwards, variants of GANs such as conditional generative adversarial networks (cGANs) were introduced for image-to-image translation tasks [14]. The pix2pix model introduced in [15] is one of the first attempts to map natural images from an input source domain to a certain target domain. Henceforth, cGANs were used for speech enhancement either utilizing raw 1D audio tracks or the 2D log-Mel time-frequency (TF) representation. For instance, speech enhancement GAN (SEGAN) is a 1D adaptation of the pix2pix model operating on 1D raw audio tracks [16]. This model was further adapted to operate on 2D TF representations, via the frequency SEGAN (FSEGAN) framework [17]. Due to the TF representations being interpreted as an implicit feature extraction, an improved speech enhancement was reported. However, both models suffer from multiple limitations. They rely mainly on pixel-wise losses, which have been reported to produce inconsistencies in the resultant local structures [15]. Additionally, both models were utilized to enhance speech tracks of fixed durations and under relatively mild noise conditions (average SNR of 10 dB).

In this work, a new adversarial approach is proposed for enhancing speech tracks based on MedGAN framework, introduced in our previous work [18] to operate on the 2D TF representation of noisy speech input. MedGAN incorporates a non-adversarial perceptual loss which penalizes the discrepancies in the feature space between the outputs and the targets. This enhances the robustness of the speech enhancement under stringent SNR conditions and challenging noise types.
Additionally, we propose a new dynamic time resolution technique to embed variable track lengths in a fixed TF representation by adapting the time overlap based on the track length. To illustrate the performance of the proposed approach, quantitative comparison is carried out against the SEGAN, FSEGAN and a classical Wiener filter approach under different noise types and SNR levels. Furthermore, the word error rate (WER) of a pre-trained automatic speech recognition (ASR) system is also evaluated.

2. DYNAMIC TIME RESOLUTION

Previously proposed architectures are designed to work on a fixed input size with respect to the pixel-dimensionality or the number of samples for FSEGAN and SEGAN, respectively. In order to accommodate this dimensionality constraint, the input track length is fixed to 1 s intervals. Accordingly, the network inference is not efficient as any arbitrary length track should be first divided into 1 s windows and then the denoising is applied sequentially on each window.

In our proposed framework, the pixel-dimensions of the input 2D TF representation is fixed to $256 \times 256$ pixels. However, the time resolution per pixel is dynamically changing based on the length of the 1D track as shown in Fig. 1. The used TF representation is computed based on short time Fourier transform (STFT) where a window function followed by FFT is applied to overlapping segments of the 1D track. To fix the frequency dimension $N_F$, the size $S$ of the overlapping segments and thus the corresponding window function should be fixed for all the input tracks. In our implementation, a hamming window of $S = 512$ samples is used to get a one-sided spectrum of $N_F = 256$ bins. To fix the time dimension $N_T = 256$ bins, the overlapping parameter $O$ of the 1D segments should be adjusted based on the input track length $L$ according to the following relation:

$$O = \left\lfloor \frac{S - L}{N_T} \right\rfloor$$

(1)

Finally, the track length $L$ is modified either by omitting or appending extra samples based on the following constraint:

$$L = N_T(S - O) + O$$

(2)

Fig. 1. Examples of variable duration tracks embedded in a fixed $256 \times 256$ TF representation.

After applying the denoising framework to the resultant TF representation, getting back to an audible 1D track is mandatory. For this we choose to use the least square inverse short time Fourier transform (LS-ISTFT) proposed in [19]. Based on this implementation an acceptable signal-to-distortion ratio (SDR) reconstruction can be achieved with an overlap of 25%. By substituting this overlap in Eq. 1, the longest track length that can be embedded in a $256 \times 256$ TF representation should not exceed 6.1 s. The LS-ISTFT requires both magnitude and phase of the TF representation for reconstruction. However, the speech phonetic patterns are mostly available in the magnitude, therefore only this component ($y_m$) is passed as input to the network. For reconstruction, the noisy phase ($y_p$) is used together with the enhanced magnitude as shown in Fig. 2.

3. METHOD

In this section, the proposed adversarial approach for speech enhancement will be described. First, a brief explanation of traditional cGANs will be outlined followed by the utilized MedGAN framework. An overview of the proposed approach is presented in Fig. 3.

3.1. Conditional Generative Adversarial Networks

In general, adversarial frameworks are a game-theoretical approach which pits multiple networks in direct competition with each other. More specifically, a cGAN framework consists of two deep convolutional neural networks (DCNNs), a generator $G$ and a discriminator $D$ [15]. The generator receives as input the 2D TF representation of the corrupted speech data $y_m$. It attempts to eliminate the intrusive background noise by outputting the synthetically enhanced TF representation $\hat{x} = G(y_m)$. The main goal of the generator is to render $\hat{x}$ to be indistinguishable from the target ground-truth clean noise signal $x$. Parallel to this process, the discriminator network is trained to directly oppose the generator. $D$ acts as a binary classifier receiving $y_m$ and either $x$ or $\hat{x}$ as inputs and classifying which of the input-pairs are synthetically generated and which are real. In other words, $G$ attempts to produce a realistically enhanced TF representation to fool $D$, while conversely $D$ constantly improves its performance to better detect the generator’s output as fake. This adversarial training setting drives both network to improve their respec-
tive performance until Nash’s equilibrium is reached. This training procedure is expressed via the following min-max optimization task over the adversarial loss function $L_{adv}$:

$$\min_{G} \max_{D} L_{adv} = \mathbb{E}_{x,y} \left[ \log D(x, y) + \mathbb{E}_{x,y} \left[ \log (1 - D(x, y)) \right] \right]$$

(3)

To further improve the output of the generator and avoid visual artifacts, an additional $L_1$ is utilized to enforce pixel-wise consistency between the generator output and the ground-truth targets [15]. The $L_1$ loss is given by

$$L_{L_1} = \mathbb{E}_{x,\hat{x}} \left[ \|x - \hat{x}\|_1 \right]$$

(4)

### 3.2. Perceptual Loss

The magnitude component of the speech TF representation is full of rich patterns directly reflecting human speech phonetics. A straightforward minimization of the pixel-wise discrepancy, via $L_1$ loss, will not enable the effective elimination of challenging speech-based noise types, e.g. cafe or food court.

To overcome this issue, we propose the utilization of the perceptual loss, based on the MedGAN framework, to regularize the generator network to produce more globally consistent results by focusing on more wider feature representations rather than individual pixels. This is achieved by utilizing the discriminator network as a trainable feature extractor to extract intermediate feature representations. The perceptual loss is then calculated as the weighted average of the mean absolute error (MAE) of the extracted feature maps:

$$L_{Percep} = \sum_{i=1}^{N} \lambda_i \| D_n(x) - D_n(\hat{x}) \|_1$$

(5)

where $D_n$ is the feature map extracted from the $n^{th}$ layer of the discriminator. $N$ and $\lambda_i$ are the total number of layers and the individual weight given to each layer, respectively.

### 3.3. Architectural Details

Based on the MedGAN framework, a CasNet generator and a patch discriminator architecture are utilized for the proposed framework [18]. CasNet concatenates three U-blocks in an end-to-end manner, whereas each U-block consists of an encoder-decoder architecture joint together via skip connections. These connections avoid the excessive loss of information due to the bottleneck layer. The output TF representations are progressively refined as they propagate through the multiple encoder-decoder pairs. The architecture of each U-block is identical to that proposed in [15]. Regarding the patch discriminator, it divides the input TF representations into smaller patches before proceeding with classifying each patch as real or fake. For the final classification score, all patch scores are averaged out. However, unlike the 70 x 70 pixel patches recommended in [15], a patch size of 16 x 16 was found to be more beneficial in this use-case.

### 4. EXPERIMENTS

The proposed speech enhancement framework is evaluated on the TIMIT dataset [20]. This dataset consists of 10 phonetically rich sentences spoken by 630 speaker with 8 different American English dialects. All the tracks are sampled at 16 kHz and the tracks durations are between 0.9 s to 7 s. The majority of the tracks are satisfying the aforementioned track length constraint in Sec. 2 with only 15 tracks exceeding the 6.1 s limit and these tracks are excluded from the dataset.

For the training procedure, three different noise types were utilized (crowded cafe, food court and moving car) from the QUT-TIMIT noise proposed in [21]. The background noise was added to the clean speech tracks in order to create a paired training set. Additionally, different SNR levels were used for each noise type (0, 5 and 10 dB). Thus, the total training dataset consists of 36,000 tracks. For validation, two different experiments were conducted. In the first experiment, the trained network was validated on a dataset of 5000 tracks utilizing the same training noise types albeit from different individuals of 2 new dialects. In the second experiment, the generalization capability of the network was investigated by validating on a dataset of 500 tracks corrupted by a new noise type, the city street noise. Both experiments were conducted using the same SNRs used in training (average SNR 5 dB).

To compare the performance of the proposed adversarial approach, quantitative comparisons were conducted against the FSEGAN, SEGAN and a traditional Weiner filter approach [17, 16, 7]. All trainable models were trained using the same hyperparameters for 50 epochs to ensure a fair comparison. Multiple metrics were used for the comparison in order to give a wider scope of interpretation for the results. The utilized metrics are the perceptual evaluation of speech quality (PSEQ) [22], the mean opinion score (MOS) prediction of the signal distortion (CSIG), the MOS prediction of background noise (CBAK) and the overall MOS prediction score (COVL) [23]. Additionally, the WER was evaluated using Google’s open-source pre-trained ASR model [24].
First a qualitative comparison on a TF representation example is illustrated to get a clear interpretation about the advantage of the proposed model. As shown in Fig. 4, the MedGAN is superior in canceling the low power components of the background noise in comparison to FSEGAN as annotated by (↓). In contrast to the MedGAN, the SEGAN model shows a clear elimination of required speech bands as annotated by (↑). This is also reflected in the quantitative comparison shown in Tables I and II. The proposed MedGAN achieves better scores over the other approaches in all the used quantitative metrics. It must also be pointed that in literature the FSEGAN authors claim a better performance in WER over the SEGAN model. However, this have not been observed in the above results. We hypothesize that this is the result of FSEGAN now having to deal with variable time resolution input TF representations, due to the utilized dynamic time resolution, which posses a challenge compared to the SEGAN.

In this work, an adversarial speech enhancement technique is introduced to operate on audio TF representations. The proposed approach involves an additional perceptual loss and a CasNet generator architecture to enhance the detailed local features of the output audio tracks. Moreover, to improve the inference efficiency, time-domain tracks with variable durations are embedded in a fixed TF representation by changing the corresponding time resolution.

Challenging speech-based noise types, e.g. cafe and food court noise, were involved in training with low average SNR (5 dB). To evaluate the generalization capability of our model, two experiments were conducted on different speakers and noise types. The proposed approach exhibits an enhanced performance in comparison to the previously introduced GAN-based and traditional model-based approaches. In future, the proposed framework will be extended to involve general audio data and enhancement tasks such as dereverberation.
7. REFERENCES

[1] M. Berouti et al., “Enhancement of speech corrupted by acoustic noise,” in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), April 1979, vol. 4, pp. 208–211.

[2] S. Kamath and P. Loizou, “A multi-band spectral subtraction method for enhancing speech corrupted by colored noise,” in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), May 2002, vol. 4, pp. 4164–4164.

[3] N. Li and P. C. Loizou, “Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction,” The Journal of the Acoustical Society of America, vol. 123, no. 3, pp. 1673–1682, 2008.

[4] D. Wang, “Time-frequency masking for speech separation and its potential for hearing aid design,” Trends in amplification, vol. 12, no. 4, pp. 332–353, 2008.

[5] J. Lim and A. Oppenheim, “All-pole modeling of degraded speech,” IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 26, no. 3, pp. 197–210, June 1978.

[6] P. Scalart and J. V. Filho, “Speech enhancement based on a priori signal to noise estimation,” in IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings (ICASSP), May 1996, vol. 2, pp. 629–632.

[7] Y. Hu and P. C. Loizou, “Speech enhancement based on wavelet thresholding the multitaper spectrum,” IEEE Transactions on Speech and Audio Processing, vol. 12, no. 1, pp. 59–67, Jan. 2004.

[8] Xugang Lu et al., “Speech enhancement based on deep denoising autoencoder,” in Interspeech, Aug. 2013, pp. 436–440.

[9] X. Feng, Y. Zhang, and J. Glass, “Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2014, pp. 1759–1763.

[10] Felix Weninger et al., “Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr,” in International Conference on Latent Variable Analysis and Signal Separation. Springer, Aug. 2015, pp. 91–99.

[11] Y. Tu et al., “A hybrid approach to combining conventional and deep learning techniques for single-channel speech enhancement and recognition,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), April 2018, pp. 2531–2535.

[12] D. Rethage, J. Pons, and X. Serra, “A wavenet for speech denoising,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), April 2018, pp. 5069–5073.

[13] Ian Goodfellow et al., “Generative adversarial nets,” in Advances in Neural Information Processing Systems, pp. 2672–2680. 2014.

[14] M. Mirza and S. Osindero, “Conditional generative adversarial nets,” CoRR, vol. abs/1411.1784, 2014.

[15] P. Isola et al., “Image-to-Image translation with conditional adversarial networks,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017, pp. 5967–5976.

[16] S. Pascual, A. Bonafonte, and J. Serrà, “SEGAN: Speech enhancement generative adversarial network,” in Interspeech, 2017, pp. 3642–3646.

[17] C. Donahue, B. Li, and R. Prabhavalkar, “Exploring speech enhancement with generative adversarial networks for robust speech recognition,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), April 2018, pp. 5024–5028.

[18] K. Armanious et al., “MedGAN: Medical image translation using GANs,” CoRR, vol. abs/1806.06397, 2018.

[19] Bin Yang, “A study of inverse short-time Fourier transform,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), March 2008, pp. 3541–3544.

[20] J. S. Garofolo et al., “TIMIT acoustic-phonetic continuous speech corpus,” Linguistic Data Consortium, 1992.

[21] D. B. Dean et al., “The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms,” in Interspeech, Sep. 2010.

[22] A. W. Rix et al., “Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs,” in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), May 2001, vol. 2, pp. 749–752.

[23] Y. Hu and P. C. Loizou, “Evaluation of objective quality measures for speech enhancement,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 16, no. 1, pp. 229–238, Jan. 2008.

[24] Google, “Cloud speech-to-text,” https://cloud.google.com/speech-to-text/