Interspeech 2018
2-6 September 2018, Hyderabad

Effectiveness of Generative Adversarial Network for Non-Audible Murmur-to-Whisper Speech Conversion

Neil Shah, Nirmesh J. Shah, and Hemant A. Patil

Speech Research Lab,
Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar, India
{neil_shah, nirmesh88_shah and hemant_patil}@daiict.ac.in

Abstract

The murmur produced by the speaker and captured by the Non-Audible Murmur (NAM)-one of the Silent Speech Interface (SSI) technique, suffers from the speech quality degradation. This is due to the lack of radiation effect at the lips and lowpass nature of the soft tissue, which attenuates the high frequency-related information. In this work, a novel method for NAM-to-Whisper (NAM2WHSP) speech conversion incorporating Generative Adversarial Network (GAN) is proposed. The GAN minimizes the distributional divergence between the whispered speech and the generated speech parameters (through adversarial optimization). The objective and subjective evaluation performed on the proposed system, justifies the ability of adversarial optimization over Maximum Likelihood (ML)-based optimization networks, such as a Deep Neural Network (DNN), in preserving and improving the speech quality and intelligibility. The adversarial optimization learns the mapping function with 54.2 % relative improvement in MOS and 29.83 % absolute reduction in % WER w.r.t. the state-of-the-art mapping techniques. Furthermore, we evaluated the proposed framework by analyzing the level of contextual information and the number of training utterances required for optimizing the network parameters, for the given task and database.

Index Terms: Non-Audible Murmur (NAM), generative adversarial network (GAN), whispered speech.

1. Introduction

Silent Speech Interface (SSI) provides a platform for producing an acoustically intelligible and sensible speech, allows processing of speech and extraction of speech-specific features, in absence of an intelligible speech [1, 2]. An SSI records the signal produced from the elements of the human speech production system, such as articulators, palate, jaw movements, neural pathways, and the brain to a certain extent. The digital representation produced through the interface can be used as an assistance to the speech-handicapped people (those suffering from the vocal tract disorder), and conveys the hidden information embedded in the silent communication systems. The interface has the great potential in generating more natural sounding, intuitive, spontaneous, and intelligible speech for the children, older, and speech-handicapped people, who require an effort in producing the speech, due to the inability of articulators' movement.

Integration of SSI with cellphone finds an important application in the modern-day communication network, where silent communication is preferred at the public places. The performance of speech processing in noisy environments can be significantly improved when speech recognition is performed on the silent speech. Here, the ambient noise would not interfere with the interface due to the sensor's robustness [1], and non-acoustically recorded speech signal. In addition, patients suffering from the vocal fold-related disorders, such as vocal fold paresis and paralysis [3, 4], etc. may not be able to produce an intelligible speech, due to the absence of partial or complete vocal fold vibrations. This inability of producing speech, severely hampers the life of the speech-handicapped person, since speech is considered as the most crucial way of communication [5]. Among the various available SSI techniques [6, 7], we focus on the Non-Audible Murmur (NAM) microphones, that are attached behind the talker's ear and can capture a very quietly uttered speech [8–11].

Study reported in [8] investigated the use of NAM, as a speech communication interface and proposed to analyze the vibrations produced within the human body, instead of analyzing the dispersed acoustic vibrations in the air. The sensor recorded parameters can then be used for speech recognition. NAM are the speech cues produced by the interactions of human speech organs, such as tongue, palate, lips, etc. and are transmitted through the soft tissue of the head [8]. This low power breathy voice produced due to the articulatory movements is termed as a silent speech or murmur [8]. Thus, the NAM waveform lacks in preserving the quality and intelligibility, due to the low-pass nature of the soft tissues and lack of radiation effect at the lips [8, 12]. Moreover, a complete description of sound radiation at the lips and diffraction above the head is quite difficult. In particular, there is no closed form expression of radiation impedance [chapter 4, pp. 130, [12]]. Thus, improving the intelligibility of the NAM speech remains a challenging task. The conversion of NAM-to-Whisper speech (NAM2WHSP), would make the speech communication possible for the patients suffering from the vocal fold disorder. The NAM2WHSP conversion improves the intelligibility and enhances the quality of the silent or breathy speech, which lacks the characteristics of vocal fold movements. Traditional NAM-to-audible speech conversion can be achieved through speech recognizer and synthesis-based approach [13] and mapping-based approaches [14–16].

In this paper, we attempt to use Generative Adversarial Network (GAN) for NAM2WHSP conversion task. To the best of authors' knowledge, this is the first attempt of its kind to apply GANs for NAM2WHSP conversion. Effectively, through supervised learning algorithm, such as GAN, we propose to learn the mapping function between the NAM speech and the whispered (audible) speech. First, we check the intelligibility of the converted audible speech, mapped through a Deep Neural Network (DNN). However, the likelihood maximization criteria, such as a Mean Square Error (MSE) loss only reduce the numerical errors between the whispered and the generated speech parameters, which may not necessarily lead to perceptually optimum speech [17, 18]. Moreover, the intelligibility

10.21437/Interspeech.2018-1565

and objective scores, suggest the need of exploiting other Deep Learning (DL) alternatives, that essentially reduces the perceptual divergence between the groundtruth and the estimated. To achieve this, we propose to exploit GAN, which optimizes an adversarial loss by minimizing the distributional divergence between the model and the data distribution. Furthermore, we also analyze the effect of varying the contextual information at the input of the network and varying the number of training examples required for optimizing the network parameters, in the NAM2WHSP conversion task. In this paper, we have shown that, when the network is trained in an adversarial framework (such as GAN), the parameters get optimized within a few iterations, with lesser number of training examples, and produces natural sounding speech with improved intelligibility, quality, and perception, than the network trained using ML-based optimization criteria (such as, DNN).

2. Proposed NAM2WHSP System

The proposed GAN-based NAM2WHSP framework is illustrated in Figure 1. First, the cepstral features are extracted from the NAM and whispered speech signal. The recording of the NAM and whispered speech is done simultaneously. Hence, we do not require any alignment before learning the mapping function. In this work, we have applied DNN and GAN-based conversion techniques to learn the mapping function. As both NAM and whispered speech are unvoiced sounds, we do not apply F_0 conversion technique. At the time of testing, we extract the cepstral features from the input NAM signal and convert it using the learned mapping function. In the end, the vocoder is employed for converting the features into the whispered speech signal.

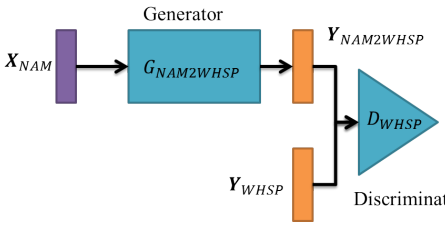

Figure 1: *Proposed schematic representation of the GAN-based NAM2WHSP conversion system.*

2.1. Generative Adversarial Network (GAN)

The recent performance improvement on modeling deep representation and learning a suitable mapping function using GAN, have shown a significant rise in speech technology-related applications, such as Voice Conversion (VC) [19, 20], speech synthesis [21] and Speech Enhancement (SE) [22–25]. GAN is a generative network that implicitly models the high-dimensional data distribution [17, 18]. Conventional generative models minimize the divergence between the model distribution $\hat{\mathcal{Y}}$ and the data distribution $\mathcal{Y}$. On the other hand, GANs are trained discriminatively to generate the samples that are indistinguishable from the actual samples y drawn from the true distribution $y \sim \mathcal{Y}$.

The GAN model comprises of a generator (G) and a discriminator (D). The G model learns a complex relationship between the samples x from the prior distribution $\mathcal{X}$ to samples $y \sim \mathcal{Y}$, and a discriminative model (D) aims to maximize the probability of correctly discriminating between the real samples y and the samples generated by the G network [18]. Both the networks are trained simultaneously in an adversarial way, forcing the G network to generate the samples belonging to $\hat{\mathcal{Y}}$, that closely follows $\mathcal{Y}$. Such a setup leaves the D network confused in discriminating the samples belonging to $\mathcal{Y}$ and $\hat{\mathcal{Y}}$. Hence, GAN produces the perceptually optimum speech by minimizing the distributional difference (i.e., divergence) between the true samples and the generated samples [18, 21]. In particular, G, i.e., G_{NAM2WHSP}, will convert NAM to Whisper speech signal, and the D, i.e., D_{WHSP} will detect whether converted whisper is perceptually similar to the true whisper. This objective can be expressed through the following optimization function:

$$\min_G \max_D \mathbb{E}_{y\sim\mathcal{Y}}[\log D_{\text{WHSP}}(y)] + \mathbb{E}_{x\sim\mathcal{X}}[1 - \log(D_{\text{WHSP}}(G_{\text{NAM2WHSP}}(x)))]. \quad (1)$$

However, the vanilla-GAN (v-GAN) architecture initially proposed in [18], may sometimes fail in optimizing the network parameters and thereby fails in preserving the speech intelligibility and the quality [25]. The adversarial training only minimizes the distributional divergence between $\mathcal{Y}$ and $\hat{\mathcal{Y}}$, but may fail in generating the mapped speech parameters corresponding to the given speech frames at the input. That is, as training proceeds, the G network produces the samples that may closely follow $\mathcal{Y}$, however, may not correspond to the given samples at the input [25]. Regularization of the adversarial loss has shown to be beneficial in learning the corresponding mapped features [22, 25]. Since our task is to learn the mapping function between the NAM and the whispered speech, we rely on the regularized adversarial objective function proposed in our earlier work [25], and can be mathematically formulated as:

$$\min_D V(D_{\text{WHSP}}) = -\mathbb{E}_{y\sim\mathcal{Y}}[\log D_{\text{WHSP}}(y)] - \mathbb{E}_{x\sim\mathcal{X}}[1 - \log(D_{\text{WHSP}}(G_{\text{NAM2WHSP}}(x)))], \quad (2)$$

$$\min_G V(G_{\text{NAM2WHSP}}) = -\mathbb{E}_{x\sim\mathcal{X}}[\log(D_{\text{WHSP}}(G_{\text{NAM2WHSP}}(x)))] + \frac{1}{2}\mathbb{E}_{y\sim\mathcal{Y},x\sim\mathcal{X}}[\log(y) - \log(G_{\text{NAM2WHSP}}(x))]^2, \quad (3)$$

where $\mathbb{E}_{y\sim\mathcal{Y}}$ denotes the expectation over all the samples y with distribution $\mathcal{Y}$ [25].

3. Experimental Results

3.1. Experimental Setup

The proposed algorithm is evaluated on the CSTR NAM TIMIT Plus corpus [26]. This corpus contains 420 newspaper texts, randomly taken from the Herald Glasgow. For our experimental analysis, 421 utterances of NAM speech and its corresponding whispered speech signals are taken from the Herald text. 400 random utterances (appx. 400k number of frames) are used to train the models and the test set comprises of the remaining 21 utterances. We train two models to evaluate the results. The first model is a DNN, with parameters optimized using the Minimum Mean Square Error (MMSE) criteria between the whispered speech and the predicted whispered speech cepstral representation. The second model is a GAN with MMSE regularization. The DNN and G network in GAN follows the identical architecture, with the three hidden layers. Having a uniform architecture helps in analyzing the advantage of using adversarial loss characteristics over the MMSE-based ML optimization. Each layer has 512 units with Rectified Linear Unit

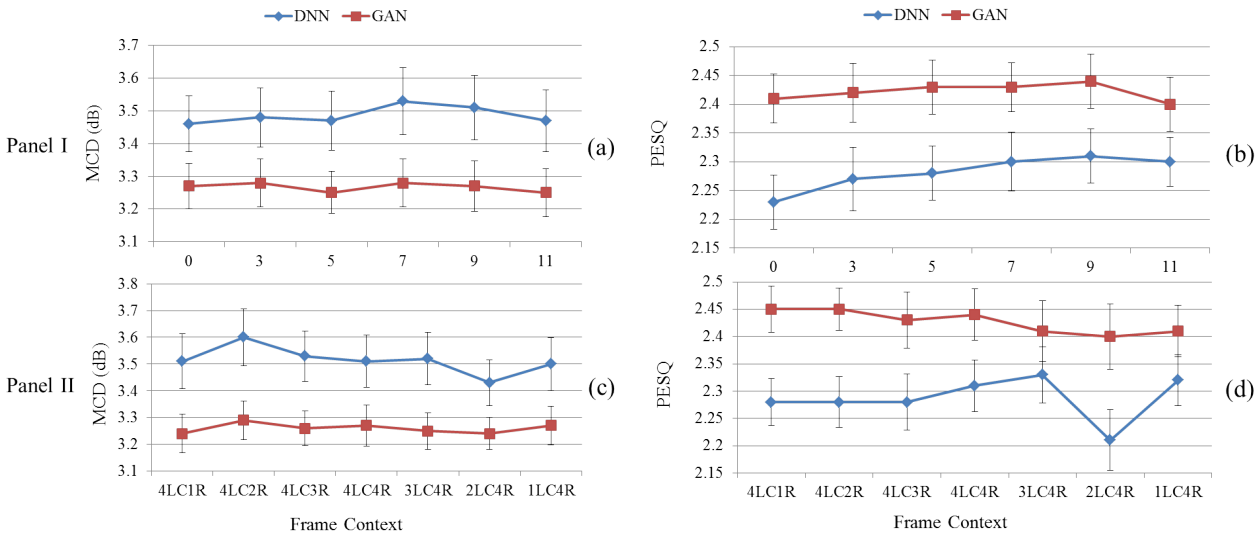

Figure 2: *MCD and PESQ analysis of different NAM2WHSP systems, Panel I: symmetric context and Panel II: asymmetric context.*

(ReLU) activation, whereas, the output layer has linear activation function. The D network in GAN also has three hidden layers, with *tanh* activation function, as suggested in [25]. The output layer has sigmoid activation that predicts the likelihood of predicted whispered speech cepstra belonging to the true distribution. Both the models are trained for 150 epochs, using an effective batch size of 1000 frames [25]. The parameters are optimized using an Adam optimization [27], with a learning rate of 0.001. We train the network by extracting the Mel Cepstral Coefficients (MCCs) from the speech signal. The original utterances of the database are downsampled from 96 kHz to 16 kHz. The *25*-dimensional (dim) MCCs (including the 0^{th} coefficient) are computed with 25 ms Hamming window and 5 ms overlap between the consecutive frames. For analysis-synthesis, we have used AHOCODER [28].

3.2. Objective Evaluation

The effectiveness of the NAM2WHSP conversion system is measured using Mel Cepstral Distortion (MCD) and Perceptual Evaluation of Speech Quality (PESQ) [29]. The traditional MCD measure is given by [30]:

$$MCD[dB] = \frac{10}{ln10}\sqrt{2\sum_{i=1}^{25}(m_i^t - m_i^c)^2} \quad , \qquad (4)$$

where m_i^t and m_i^c are the i^{th} MCCs of the whispered and converted whispered speech cepstral features. The PESQ measure evaluates the quality of the speech [29]. The lower MCD and higher PESQ measure signifies the improved system performance using the proposed system.

3.2.1. Effect of contextual information in NAM2WHSP system

Since speech is a sequential data, extracting the contextual features from the speech, captures the local features (including coarticulation) and preserves the crucial harmonics [31, 32]. In speech perception, it has been shown that the surrounding acoustic context, impacts the human perception [33–35]. Recently, researchers from the neuroscience of speech perception have tried to identify the underlying representations in the primary and secondary auditory cortex, and have examined the information modulated by varying the context [36]. Motivated from this, we analyze the effect of varying context size window (with window length 0, 3, 5, 7, 9, and 11) at the network input. Here, we extract the different symmetric contextual NAM speech features. These networks are trained to predict the *25*-dim MCCs of the whispered speech. Out of the total 400 training utterances, 350 random utterances are used for training the models and remaining 50 utterances are used for validation. Once the network is trained, the epoch with the least MSE on the validation set is selected for the testing purpose. Figure 2 clearly demonstrates the effect of context window length variations on the performance of the developed systems in terms of both the objective scores. Panel I (Figure 2 (a) and (b)) shows the MCD and PESQ scores, for the systems developed on the different number of symmetric context frames. The effectiveness of the GAN-based NAM2WHSP system over DNN is clearly observed in both the objective test. From both the PESQ and MCD analysis, it can be seen that system with 9 window context (i.e., four frames on the left and four on the right) yields the highest PESQ and the almost second least MCD score. Hence, the 9 frames context window system is selected for our further analysis.

Inspired by the study reported in [32], we also analyze the importance of training models by taking an asymmetric contextual frames as an input to the network (Panel II in Figure 2). No significant variations could be observed in terms of MCD scores for GAN-based systems (as shown in Figure 2 (c)), whereas significant improvement in the performance of the GAN-based system over DNN-based system, is observed in terms of PESQ score (as shown in Figure 2 (d)) (notably 4LC1R). From Figure 2 (c) and (d), it can be observed that, when more weight is given to the left phoneme in an asymmetric combination, the GAN significantly outperforms the other asymmetric combinations.

3.2.2. Analysis w.r.t. amount of training data

We also analyze the models on varying the size of training data. Figure 3 (a) and (b) shows that the GAN predicts better MCD and PESQ score, with an increased number of training utterances. Initially, DNN dominates the GAN, however, as the number of training utterance increases, the objective scores predicted by the DNN deteriorates significantly. This may be due to the fact that, initially with a very few training utterances, the

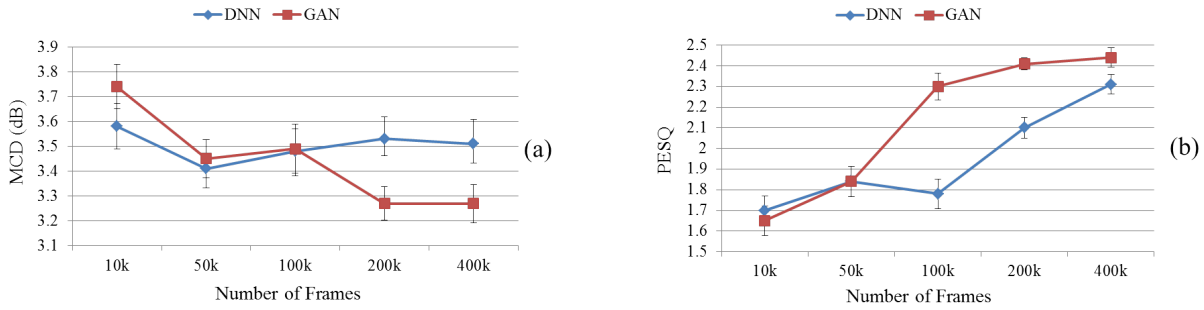

Figure 3: *(a) MCD and (b) PESQ analysis of the various developed NAM2WHSP systems w.r.t. the amount of available training data.*

discriminator in the GAN is very confident about its decision of rejecting the generated cepstrum. Moreover, with generator exposed to a very few training utterances, though indirectly, may not be able to sufficiently fool the discriminator. However, with the increased exposure of training utterances, the adversarial training forces the generator to produce the cepstrum that approximately follows the data distribution, and successfully confuses the discriminator. It has to be noted that the poor performance measure exhibited by DNN, may be due to the absence of such an adversarial nature in training (ML-based optimization).

3.3. Subjective Evaluation

The key objective of the proposed work is to extract the message recorded via NAM microphone. Hence, we focus on the various intelligibility tests for the subjective evaluation. We consider two tests, namely, Word Error Rate (WER) and Mean Opinion Scores (MOS) test for the intelligibility [37]. Total 31 listeners (21 males and 10 females with age between 18 to 30 years) took part in all the subjective tests. We used high-quality Sennheiser headphones for the subjective evaluations.

Table 1: *% WER analysis for the developed NAM2WHSP systems*

	NAM	DNN	GAN (Proposed)	WHSP
WER (%)	69.3	65.5	35.67	5.74
Number of Replays	3.39	3.2	2.2	1.43

In WER test, we asked subjects to transcribe eight randomly played utterances from the original NAM and whispered speech (WHSP), and predicted whispered speech using the DNN and the GAN-based systems. Moreover, we asked the subjects not to replay any utterance more than four times (in order to avoid cognitive related *bias* in hearing and subjective judgment) during their transcription. Based on their submission, % WER is calculated as [38]:

$$WER(\%) = \frac{I + D + S}{T} \times 100, \tag{5}$$

where I, D and S represents the number of insertions, deletions, and substitutions, respectively, and T is the total number of words in a given utterance. From Table 1, it can be observed that the % WER of the NAM is very high. However, for the whispered speech, % WER is less. Our proposed GAN-based system obtained an absolute reduction of 29.83 in % WER compared to the DNN-based NAM2WHSP system. In addition, the number of replays required for the GAN-based system is less than the one required by the DNN-based system. We used the same DNN architecture along with the same hyper parameters, which is used in the generator.

Five utterances from each system were selected to evaluate the MOS for the intelligibility test. Subjects were asked to rate each randomly played utterance based on the intelligibility from the different systems on a five-point scale (1= not at all intelligible; 2= hardly one or two words are intelligible; 3= half of the message is intelligible; 4= mostly intelligible, and 5= completely intelligible). The analysis of MOS along with the 95 % confidence interval is shown in Figure 4. There is a clear 54.2 % of relative improvement in the MOS obtained using GAN-based system compared to the DNN-based NAM2WHSP system. Lower MOS for the NAM signal clearly indicates the less intelligibility of the NAM signal. This is primarily due to the lack of radiation effect at the lips and lowpass nature of the soft tissue. Hence, the objective of extracting the linguistic message from the NAM signal using the GAN-based approach is indeed achieved to a certain extent.

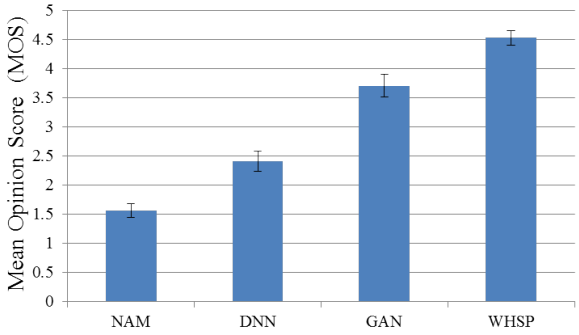

Figure 4: *MOS analysis for intelligibility of various systems along with 95 % confidence interval.*

4. Summary and Conclusions

In this work, we proposed the novel GAN-based NAM2WHSP conversion system. The GAN-based training achieves 54.2 % relative improvement in MOS and 29.83 % absolute reduction in WER w.r.t. state-of-the-art DNN-based systems, due to its adversarial nature. The objective as well as the subjective evaluation indicates the importance of minimizing the distributional divergence in learning the mapping function and preserving the speech quality and achieving large intelligibility gains over the traditional ML-based optimization techniques. In addition, we identify the impact of symmetric as well as asymmetric contextual frames, and the number of training utterances required for optimizing the network parameters. We plan to evaluate the effectiveness of the proposed font-end DNN/GAN-based NAM2WHSP conversion system via stat-of-the-art whisper speech recognition system.

5. Acknowledgments

Authors thank authorities of DA-IICT, Gandhinagar and MeitY, Govt. of India for their kind support.

6. References

[1] B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg, "Silent speech interfaces," *Speech Communication*, vol. 52, no. 4, pp. 270–287, 2010.

[2] T. Schultz and M. Wand, "Modeling coarticulation in EMG-based continuous speech recognition," *Speech Communication*, vol. 52, no. 4, pp. 341–353, 2010.

[3] A. D. Rubin and R. T. Sataloff, "Vocal fold paresis and paralysis," *Otolaryngologic Clinics of North America*, vol. 40, no. 5, pp. 1109–1131, 2007.

[4] L. Sulica, "Vocal fold paresis: an evolving clinical concept," *Current Otorhinolaryngology Reports*, vol. 1, no. 3, pp. 158–162, 2013.

[5] J. A. Mattiske, J. M. Oates, and K. M. Greenwood, "Vocal problems among teachers: a review of prevalence, causes, prevention, and treatment," *Journal of Voice*, vol. 12, no. 4, pp. 489–499, 1998.

[6] S.-C. S. Jou, T. Schultz, and A. Waibel, "Adaptation for soft whisper recognition using a throat microphone." in *INTERSPEECH*, Jeju Island, Korea, 2004, pp. 1493–1496.

[7] S.-C. Jou, T. Schultz, M. Walliczek, F. Kraft, and A. Waibel, "Towards continuous speech recognition using surface electromyography," in *INTERSPEECH*, Pittsburgh, USA, 2006, pp. 573–576.

[8] Y. Nakajima, H. Kashioka, K. Shikano, and N. Campbell, "Non-audible murmur recognition input interface using stethoscopic microphone attached to the skin," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Hong Kong, 2003, pp. 704–708.

[9] T. Toda, K. Nakamura, H. Sekimoto, and K. Shikano, "Voice conversion for various types of body transmitted speech," in *ICASSP*, Taipei, Taiwan, 2009, pp. 3601–3604.

[10] Y. Tajiri, H. Kameoka, and T. Toda, "A noise suppression method for body-conducted soft speech based on non-negative tensor factorization of air-and body-conducted signals," in *ICASSP*, New Orleans, USA, 2017, pp. 4960–4964.

[11] P. Heracleous, T. Kaino, H. Saruwatari, and K. Shikano, "Unvoiced speech recognition using tissue-conductive acoustic sensor," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, no. 1, pp. 001–011, 2006.

[12] T. F. Quatieri, *Discrete Time Speech Signal Processing: Principles and Practice*. Pearson Education India, 1^{st} (Eds.), 2006.

[13] T. Hueber, E.-L. Benaroya, G. Chollet, B. Denby, G. Dreyfus, and M. Stone, "Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips," *Speech Communication*, vol. 52, no. 4, pp. 288–300, 2010.

[14] T. Toda and K. Shikano, "NAM-to-speech conversion with Gaussian mixture models," in *INTERSPEECH*, Lisbon, Portugal, 2005, pp. 1957–1960.

[15] V.-A. Tran, G. Bailly, H. Lœvenbruck, and T. Toda, "Multimodal HMM-based NAM-to-speech conversion," in *INTERSPEECH*, Brighton, United Kingdom (UK), 2009, pp. 656–659.

[16] V. A. Tran, G. Bailly, H. Lœvenbruck, and T. Toda, "Improvement to a NAM-captured whisper-to-speech system," *Speech Communication*, vol. 52, no. 4, pp. 314–326, 2010.

[17] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 53–65, 2018.

[18] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems (NIPS)*, Montral, Canada, 2014, pp. 2672–2680.

[19] T. Kaneko, H. Kameoka, K. Hiramatsu, and K. Kashino, "Sequence-to-sequence voice conversion with similarity metric learned using generative adversarial networks," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 1283–1287.

[20] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from unaligned corpora using variational autoencoding Wasserstein generative adversarial networks," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 3364–3368.

[21] Y. Saito, S. Takamichi, H. Saruwatari, Y. Saito, S. Takamichi, and H. Saruwatari, "Statistical parametric speech synthesis incorporating generative adversarial networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 26, no. 1, pp. 84–96, 2018.

[22] D. Michelsanti and Z.-H. Tan, "Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 3642–3646.

[23] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: speech enhancement generative adversarial network," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 3642–3646.

[24] T. Kaneko, S. Takaki, H. Kameoka, and J. Yamagishi, "Generative adversarial network-based postfilter for STFT spectrograms," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 3389–3393.

[25] M. H. Soni, N. Shah, and H. A. Patil, "Time-frequency masking-based speech enhancement using generative adversarial network," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, 2018, pp. 5039–5043.

[26] "Publicly available: The CSTR NAM TIMIT Plus Corpus," URL: homepages.inf.ed.ac.uk/jyamagis/release/CSTR-NAM-TIMIT-Plus-ver0.81.tar.gz, {Last Accessed: March 15, 2018}.

[27] D. Kingma and J. Ba, "ADAM: A method for stochastic optimization," in *International Confernece on Learning Representation (ICLR)*, San Diego, USA, 2015, pp. 1–15.

[28] D. Erro, I. Sainz, E. Navas, and I. Hernáez, "Improved HNM-based vocoder for statistical synthesizers," in *INTERSPEECH*, Florence, Italy, 2011, pp. 1809–1812.

[29] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 16, no. 1, pp. 229–238, 2008.

[30] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Trans. on Audio, Speech and Lang. Process. (TASLP)*, vol. 15, no. 8, pp. 2222–2235, 2007.

[31] H. A. Bourlard and N. Morgan, *Connectionist Speech Recognition: A Hybrid Approach*. Springer Science & Business Media, First Edition, 2012.

[32] N. J. Shah, M. Zaki, and H. A. Patil, "Influence of various asymmetrical contextual factors for TTS in a low resource language," in *International Conference on Asian Language Processing (IALP)*, Singapore, 2014, pp. 107–110.

[33] M. H. Davis, M. A. Ford, F. Kherif, and I. S. Johnsrude, "Does semantic context benefit speech understanding through top–down processes? evidence from time-resolved sparse fMRI," *Journal of Cognitive Neuroscience*, vol. 23, no. 12, pp. 3914–3932, 2011.

[34] L. L. Holt and A. J. Lotto, "Speech perception within an auditory cognitive science framework," *Current Directions in Psychological Science*, vol. 17, no. 1, pp. 42–46, 2008.

[35] M. Chait, D. Poeppel, A. De Cheveigné, and J. Z. Simon, "Processing asymmetry of transitions between order and disorder in human auditory cortex," *Journal of Neuroscience*, vol. 27, no. 19, pp. 5207–5214, 2007.

[36] M. K. Leonard and E. F. Chang, "Dynamic speech representations in the human temporal lobe," *Trends in Cognitive Sciences*, vol. 18, no. 9, pp. 472–479, 2014.

[37] I. Rec, "P. 85. a method for subjective performance assessment of the quality of speech voice output devices," *International Telecommunication Union (ITU), Geneva.*, Available Online:{https://www.itu.int/rec/T-REC-P.85-199406-I/en} {Last Accessed: March 15, 2018}.

[38] L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. PTR Prentice Hall, 1^{st} Eds., 1993.