INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Using a Manifold Vocoder for Spectral Voice and Style Conversion

Tuan Dinh[1], *Alexander Kain*[1], *Kris Tjaden*[2]

[1]Oregon Health & Science University
[2]University of New York at Buffalo

dintu@ohsu.edu, kaina@ohsu.edu, tjaden@buffalo.edu

Abstract

We propose a new type of spectral feature that is both compact and interpolable, and thus ideally suited for regression approaches that involve averaging. The feature is realized by means of a speaker-independent variational autoencoder (VAE), which learns a latent space based on the low-dimensional manifold of high-resolution speech spectra. In vocoding experiments, we showed that using a 12-dimensional VAE feature (VAE-12) resulted in significantly better perceived speech quality compared to a 12-dimensional MCEP feature. In voice conversion experiments, using VAE-12 resulted in significantly better perceived speech quality as compared to 40-dimensional MCEPs, with similar speaker accuracy. In habitual to clear style conversion experiments, we significantly improved the speech intelligibility for one of three speakers, using a custom skip-connection deep neural network, with the average keyword recall accuracy increasing from 24% to 46%.

Index Terms: intelligibility, voice conversion, style conversion, speech coding, variational autoencoder

1. Introduction

Understanding speech can be a challenge for hearing-impaired listeners, and also for normal-hearing listeners in adverse environments. In an effort to increase the intelligibility of speech, researchers used noise-suppression algorithms [1] and de-reverberation methods [2] to eliminate background noise and reverberation. Other approaches included dynamic range compression [3, 4], peak-to-rms reduction [5], and optimizations based on a speech intelligibility index [6] or glimpse proportion measure [7, 8], typically resulting in no or small improvements.

In related research, it was demonstrated that speech intelligibility improves if a speaker changes their *habitual* speaking style to a *clear* speaking style [9, 10, 11]. Typically, clear speech is highly-articulated, with a slower speaking rate, and more frequent pauses; the exact strategy varies from speaker-to-speaker. Previously, we used hybridization experiments to establish that speech intelligibility of habitual speech can be increased when certain acoustic features from parallel clear speech are incorporated [12, 13]. This suggests that it should be possible to automatically increase the intelligibility of speech by learning a mapping between habitual and clear features, or *style conversion*. However, our previous mapping experiments only showed very modest improvements, and were conducted only on vowels [14]. More recently, a mixed-filtering approach [15], which isolated, then boosted important frequency regions in habitual speech, resulted in an objective (but not subjective) improvement of speech intelligibility and a subjective improvement of speech quality.

This material is based upon work supported by the National Institutes of Health under Grant R01DC004689.

In the closely-related area of *voice conversion* (VC), we previously employed artificial neural networks to map 40-dimensional mel-cepstral coefficients (MCEP), a commonly-used short-term spectral feature in speech processing, between source and target speakers. In an ongoing effort to improve the efficacy of spectral conversion algorithms, we have developed a new type of feature that is highly efficient because it is represented by a lower-dimensional manifold that aims to cover multi-speaker speech data. Moreover, the feature is interpolable, which ensures that even when two or more parameter vectors are averaged (e. g. as part of the mapping procedure), the result remains near the manifold representing possible speech. The feature is realized by means of a speaker-independent variational autoencoder (VAE), which learns the low-dimensional manifold of speech spectra, and represents this in a latent space that allows perceptually meaningful interpolation; the visual equivalent would be morphing one digit to another in such a way that intermediate steps appear similar to both starting and ending shapes, as opposed to simply cross-fading two images.

In Section 2, we first describe the structure and training of the VAE used in the subsequent experiments. We show that using a 12-dimensional VAE feature (VAE-12) resulted in significantly better perceived speech quality as compared to a 12-dimensional MCEP feature in vocoding experiments. In Section 3, we report on VC experiments in which VAE-12 had significantly better perceived speech quality as compared to 40-dimensional MCEPs, with similar speaker accuracy. Finally, in Section 4, we first investigate which speakers benefit from using clear spectra in place of habitual spectra, using a hybridization approach [12, 13]. Next, we report on style conversion experiments in which we significantly improved the speech intelligibility of one of three selected speakers, using VAE-12 features in combination with a custom skip-connection deep neural network.

2. Experiment: Manifold Vocoder

We propose a new type of feature that is both compact and interpolable, and thus ideally suited for regression approaches that involve averaging. Both compactness and interpolability is realized through projection of high-dimensional acoustic features onto a lower-dimensional manifold that is learned from a large multi-speaker database of speech data. Thus, the features are specialized to only model acoustic events related to speech, as opposed to music or other sources. Moreover, interpolability ensures that even when two or more parameter vectors are averaged, the result remains near the manifold of possible speech; this property does not hold for MCEPs, linear predictive coefficients, or the log-magnitude discrete-time Fourier spectrum. Manifold learning is implemented with the use of a variational autoencoder.

http://dx.doi.org/10.21437/Interspeech.2019-1176

2.1. Variational Autoencoder

The variational autoencoder (VAE) is a latent variable generative model, which combines variational inference and deep learning [16]. The latent variable generative model $p_\theta(\mathbf{x}|\mathbf{z})$, also called the decoder, is a deep neural network (DNN) with parameters θ. The inference model $q_\phi(\mathbf{z}|\mathbf{x})$, also called the encoder, is represented by another DNN with parameters ϕ. The latent variable $\mathbf{z}$ is a compact representation of the observation $\mathbf{x}$, generated by the encoder mapping the input space into its corresponding latent space. In our paper, the VAE encoder predicts the mean $\mu_\mathbf{z}$ and log-variance $\log \sigma_\mathbf{z}^2$ of the posterior distribution $q_\phi(\mathbf{z}|\mathbf{x})$ from a 39-dimensional input vector. The decoder predicts the observation $\hat{\mathbf{x}}$ from samples of $\mathbf{z}$. We learn the parameters θ and ϕ by maximizing the variational lower bound $\mathcal{L}(\theta, \phi; \mathbf{x})$ given by

$$\mathcal{L}(\theta, \phi; \mathbf{x}) = \mathop{\mathbb{E}}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log(p_\theta(\mathbf{z}|\mathbf{x}))] - D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x})|p_\theta(\mathbf{z}|\mathbf{x}))$$

where D_{KL} denotes the Kullback-Leibler divergence.

Often, the inference model $q_\phi(\mathbf{z}|\mathbf{x})$ is parameterized using a diagonal Gaussian distribution $\mathcal{N}(\mathbf{z}; \mu_\mathbf{z}, \sigma_\mathbf{z}^2\mathbf{I})$. The prior is modeled as an isotopic parameterless Gaussian distribution $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I})$. To yield a differentiable network after sampling, we use the common technique of re-parameterizing the random variable $\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})$ as a deterministic variable $\mathbf{z} = \boldsymbol{\mu}_\mathbf{z} + \boldsymbol{\sigma}_\mathbf{z} \odot \boldsymbol{\epsilon}$, where $\odot$ denotes an element-wise product, and vector $\boldsymbol{\epsilon}$ is sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Recent studies [17, 18, 19] showed the efficacy and interpolability of VAE-based latent representation in modeling and transforming speech.

2.2. Method

For initial analysis and final synthesis, we used the WORLD vocoder [20, 21]. Our process was inserted as additional steps after analysis and before synthesis. Specifically, we first calculated the 40-dimensional MCEP (MCEP-40) from the inverse Fourier transform of the 512-point mel-warped log spectrum. (Resynthesis from MCEP-40 resulted in no perceivable degradation as compared to the vocoder using the high-resolution spectrum.) Next, we subtracted the mean, calculated on the whole dataset, with the goal of reducing the channel effect of a particular dataset (cepstral mean subtraction); we also excluded the zeroth coefficient, representing energy. The resulting 39-dimensional vector was then encoded as a 12-dimensional latent representation by the VAE, and then immediately decoded. We re-added the zeroth-coefficient unmodified. Finally, we calculated a new high-resolution spectrum from the resulting MCEP, and synthesized a new speech waveform, using the original fundamental frequency and aperiodicity information.

The VAE encoder consisted of three fully-connected layers with 256 nodes each and a 12-dimensional Gaussian parametric layer modeling $\mathbf{z}$. No activation function was applied to the Gaussian parametric layer. For other layers, we used rectified linear units. The decoder is identical except for the Gaussian layer. We used $\mu_\mathbf{z}$ as our compact representation of $\mathbf{x}$. To train the VAE, we used the TIMIT [22] dataset. Of the 630 available speakers we selected all 462 speakers designated for training and all 144 designated for validation. As is convention, we eliminated the spoken dialect samples (SA sentences) for all speakers. We trained with the Adam optimizer [23], a mini-batch size of 256, and early stopping.

In addition to our proposed system with 12-dimensional VAE features (VAE-12), we also implemented two other systems for comparison. The MCEP-12 system used 12^{th}-order MCEP to represent the spectrum; here we chose the order/dimensionality to be the same as the VAE-12 system. The LSF-20 system used 20^{th}-order linear predictive coefficients converted to line spectral frequencies (LSF) to represent the spectrum, calculated using the autocorrelation method derived from the inverse Fourier transform of the squared spectrum. The LSF order was chosen to produce an expected log-spectral distortion (LSD) approximately similar to that of VAE-12.

2.3. Evaluation: Speech quality

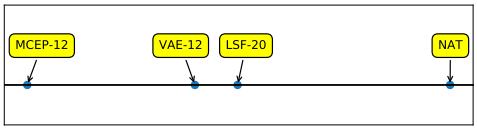
To evaluate vocoding quality, we used the voice conversion challenge (VCC) 2016 database [24], which features 5 male and 5 female speakers. Each speaker has 162 parallel utterances. We arbitrarily selected two female (SF1, TF1) and two male speakers (SM2, TM1). Using these four speakers and all available sentences, the mean (and standard deviation in parentheses) LSD in dB produced by the three vocoding systems were as follows: VAE-12 8.0 (3.0), MCEP-12: 9.18 (3.0), and LSF-20: 8.7 (3.4). Objectively, it appeared that the three systems are roughly comparable. However, it is known that the LSD measure is a poor predictor of human perception.

To evaluate vocoding quality perceptually, we selected the comparative mean opinion score (CMOS) testing approach to compare the speech quality of the three vocoding systems and natural speech (NAT). At each trial, participants listened to samples A and B in sequence and were then asked: "Is A more natural than B?" Responses were selected from a 5-point scale that consisted of "definitely better" (+2), "better" (+1), "same" (0), "worse" (−1), and "definitely worse" (−2). The test involved 4 speakers, 32 sentences, and 4 systems and thus 6 condition pairs, resulting in $4 \times 32 \times 6 = 768$ unique trials. The perceived loudness differences between these stimuli were minimized using a root-mean-square A-weighted (RMSA) measure [25]. We limited each listener to hear each unique sentence once (presentation order was randomized); therefore we needed $768 \div 32 = 24$ listeners to cover all trials. This and subsequent experiments were conducted on Amazon Mechanical Turk (AMT); we required listeners to have an approval rate $\geq 90\%$ and to live in the U. S. Table 1 shows the pair-wise relative quality of the systems (after an appropriate transformation handling the random presentation order of A and B).

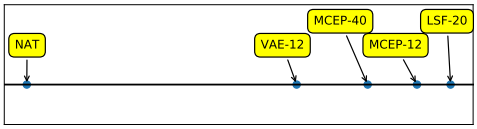
To approximate the ordering between all systems, we projected the non-negative pair-wise relative quality matrix to a single dimension, using multiple dimensional scaling (MDS), a dimensionality reduction technique that attempts to preserve the pair-wise distances of data points. Figure 1a shows the result. All synthetic systems were statistically significantly different from NAT; the difference between MCEP-12 and LSF-20, as well as the difference between MCEP-12 and VAE-12 were also significant; the latter indicates that VAE-12 was able to code the speech spectrum more efficiently.

Table 1: *Relative quality between original and vocoded stimuli. Positive values indicate A is better than B. Results marked with an asterisk are significantly different ($p < 0.001$, adjusted critical $\alpha = 0.008$) as compared to 0 (representing no preference) in a 1-sample t-test.*

A \ B	LSF-20	MCEP-12	VAE-12
NAT	+0.77*	+1.34*	+1.02*
LSF-20		+1.08*	−0.04
MCEP-12			−0.44*



(a) *Vocoding quality*



(b) *Voice conversion quality*

Figure 1: *Multiple dimensional scaling results*

3. Experiment: Voice Conversion

To further study the efficacy of our proposed feature, we performed a VC experiment using the spectral features of the four systems: MCEP-40 (a popular feature for VC), VAE-12 (our proposed method), MCEP-12 (dimension-matched comparison to VAE), and LSF-20 (another classic VC feature). We selected four source/target speaker pairs, two intra-gender and two inter-gender, from the VCC dataset: SM2→TM1 (M2M), SM2→TF1 (M2F), SF1→TF1 (F2F), SF1→TM1 (F2M). We divided the available 163 sentences into 100 training, 30 validation, and 32 test sentences.

During training, we first aligned all sentences of a given source and target speakers using dynamic time warping (DTW) on 32^{nd}-order log filter bank features. Next, we analyzed sentences with all the systems. Finally, we trained a spectral mapping from source to target features for each system. The mapping was implemented by a deep neural network (DNN) with four hidden layers of 512 nodes each. For each layer we used batch normalization, parametric relu [26], and dropout (at a rate of 20%). For the input vector, we added context by concatenating the current frame with the five preceding and the 5 following frames. We normalized the input and outputs of the network via standard scaling. Similar to training the VAE, we used the Adam optimizer, a mini-batch size of 256, and early stopping, for this and subsequent mapping experiments.

During conversion of the test sentences, we first analyzed the source with the vocoder, then computed the desired spectral feature, and mapped it. In order to measure spectral mapping performance in isolation, we created stimuli from the mapped and aligned (to the target) spectral features, and the unmodified target energy, F0, and aperiodicity information. For the LSF-20 system, when necessary, we sorted the mapped features per frame to satisfy the required monotonicity of LSFs. Finally, we minimized loudness differences using the RMSA measure.

3.1. Evaluation: Speaker accuracy and Speech quality

We used a second CMOS test to evaluate the accuracy of having converted the source speaker to sound like the target speaker. In this test, listeners heard stimuli A and B with different linguistic content, and were then asked "is B spoken by the same speaker as A?" Listeners responded using a 5-point scale comprised of "definitely same" (+2), "same" (+1), "unsure" (0), "different" (−1), and "definitely different" (−2). One stimulus was a converted sample and the other was an unmodified (NAT) reference

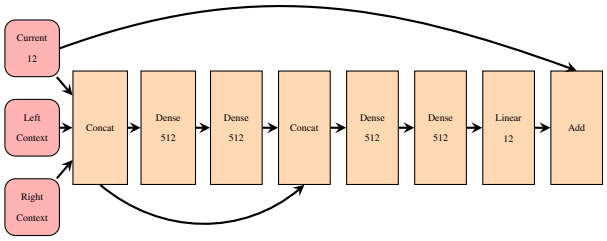


Figure 2: *DNN architecture with skip connection*

speaker. Half of all stimuli pairs were created with the reference speaker identical to the target speaker of the conversion (the "same" condition); the other half were created with the reference speaker being of the same gender, but not identical to the target speaker of the conversion (the "different" condition). Our test involved 32 sentences × 4 systems × 4 conversions × 2 conditions ("same" or a "different") = 1,024 unique trials. The experiment was conducted on AMT (with the same requirements as in Section 2.3) with 1024 ÷ 32 = 32 participants to cover all trials. Table 2 shows that VAE-12 had the best average conversion performance (0.7). We used a two-tailed t-test to compare VAE-12 to other systems and found that it was statistically significantly different ($p < 0.05$) to LSF-20.

We used a third CMOS test to compare the speech quality of the four mapping systems and the NAT condition ("is A better than B?"), using a 5-point scale comprised of "definitely better" (+2), "better" (+1), "unsure" (0), "worse" (−1), and "definitely worse" (−2). The test involved 32 sentences × 4 systems (including all 4 conversions for each system) × 10 condition pairs, resulting in 1,280 unique trials. We conducted the listening test on AMT with 1,280 ÷ 32 = 40 listeners, with each listener hearing 32 unique sentence materials. The results in Table 4 show that VAE-12 outperformed MCEP-40 and LSF-20. The multidimensional scaling shown in Figure 1b shows that VAE-12 was the closest to NAT. As a result, we used VAE-12 in the following style conversion experiment.

4. Experiment: Style Conversion

We applied the voice conversion method to convert between two *styles* of a speaker; specifically, we aimed to convert the spectral aspects of a habitually-spoken sentence (HAB) to those of a clearly-spoken sentence (CLR), in an effort to improve the speech intelligibility of the former.

We used a database with 78 speakers consisting of control speakers (CS, N=32), speakers with multiple sclerosis (MS, N=30), and speakers with Parkinson's disease (PD, N=16) [27, 28]. All read the same 25 Harvard sentences in habitual and clear conditions (loud, slow, and fast conditions were also available). We started by selecting the speakers who were previ-

Table 2: *Speaker accuracy for the same condition*

pair\system	LSF-20	MCEP-12	MCEP-40	VAE-12
F2M	0.6 (1.2)	0.18 (1.7)	0.0 (1.4)	0.47 (1.14)
M2F	0.4 (1.2)	0.8 (1.2)	1.0 (1.17)	0.9 (1.4)
F2F	0.5 (1.17)	1.0 (1.16)	0.6 (1.19)	0.7 (1.4)
M2M	-0.3 (1.3)	-0.5 (1.1)	0.5 (1.2)	0.8 (1.2)
average	0.3 (1.27)	0.4 (1.4)	0.5 (1.3)	**0.7** (1.4)

Table 3: *Average keyword accuracy measuring speech intelligibility. Results marked with an asterisk indicated that the hybrid condition was significantly different ($p < 0.05$) from the vocoded HAB condition, using a two-tailed t-test.*

	MSF7	MSF15	PDF3	PDF7	PDM6	CSM4	CSM8	CSM7	CSM6	CSF8	CSF12
vocoded HAB	68	45	5	13	30	26	42	35	30	26	18
hybrid	59	37	7	22*	55*	30	43	53*	39*	25	24
vocoded CLR	80*	53	11*	25*	56*	74*	49	65*	60*	44*	53*

ously shown to have the highest intelligibility difference between clear and habitual in their category, by using a minimum threshold of 20% absolute difference; this resulted in 11 speakers (6 CS, 2 MS, and 3 PD).

4.1. Hybridization

We first needed to establish which speakers benefit from using the CLR spectrum in place of the HAB spectrum (via a hybridization approach), as speakers use different strategies to produce CLR speech. To this end, we measured the intelligibility of hybridized stimuli [12, 13], compared with a purely-vocoded HAB condition, and a purely-vocoded CLR condition. The hybridized stimuli were created by combining the CLR spectrum, aligned to the HAB style, with HAB F0, and HAB aperiodicity information, using the WORLD vocoder. We minimized the loudness differences of stimuli by using an RMSA measure. Finally, each utterance was mixed with babble noise at 0 dB SNR to avoid saturation effects.

The speech intelligibility test design consisted of 25 sentences $\times$ 11 speakers $\times$ 3 conditions = 825 unique trials. We performed the test on AMT, wherein 66 participants listened to 25 Harvard utterances, which contain five keywords each. Listeners typed out each sentence as best as they could; their responses were subsequently manually scored. We then calculated the average number of keywords correctly identified. Table 3 shows the average keyword accuracy. We observed that spectral hybridization led to statistically significant improvements in speakers PDF7, PDM6, CSM7, and CSM6, but also resulted in degradations for MSF7 and MSF15.

4.2. Mapping

We evaluated the efficacy of our proposed VAE-12 system for the purpose of mapping HAB style to CLR style and thus improving speech intelligibility for speakers that have shown to benefit from the CLR spectrum. We used the top three speakers PDF7, PDM6, and CSM7 that showed the most benefit in the hybridization experiment. We aligned each HAB utterance to its parallel CLR utterance of the same speaker using DTW on 32^{nd}-order log filter-bank features. Then, we trained speaker-dependent mappings from HAB VAE-12 to CLR VAE-12, where we used two different DNN structures. The first structure was a typical feedforward network used in the previous voice conversion experiment (also called DNN-mapping VAE). For the second structure (also called Skip-mapping VAE), we introduced skip connections [29], as shown in Figure 2.

We created conversion stimuli consisting of the mapped VAE-12, and F0 and aperiodicity information from the original HAB speech. To create the 25 conversion sentences, we used a leave-one-out approach. Otherwise, network configurations and training parameters were identical to the VC experiment.

Table 4: *Relative quality between vocoded target and mapping, results marked with an asterisk are significantly different ($p < 0.001$, adjusted critical $\alpha = 0.005$) as compared to zero (no preference) in a one-sample t-test.*

A \ B	LSF-20	MCEP-12	MCEP-40	VAE-12
NAT	1.8*	1.8*	1.67*	1.7*
LSF-20		0.0	-0.54*	-1.08*
MCEP-12			-0.27*	-0.5*
MCEP-40				-0.5*

4.3. Evaluation: Speech Intelligibility

To evaluate speech intelligibility, we designed a test consisting of 25 sentences $\times$ 3 speakers $\times$ 5 conditions (2 purely vocoded, 1 hybrid, 2 mappings) = 375 unique trails. The test was conducted similarly to the previous one in Section 4.1, except 30 listeners participated. The hybrid stimuli show an upper bound (or "oracle" mapping) on the intelligibility for the VAE-mapping. Table 5 shows average keyword accuracy. We observed that the VAE-mapping using a custom DNN with skip connection led to a statistically significant improvement for speaker PDM6, but no significant differences in other cases, using a two-tailed *t*-test.

5. Conclusion

We proposed a compact and interpolable feature for spectral regression, implemented by a speaker-independent VAE. In a vocoding experiment, we showed that using VAE-12 achieved significantly better perceived speech quality compared to a MCEP-12 feature. In a voice conversion experiment, we showed that mapping VAE-12 resulted in significantly better perceived speech quality compared to a MCEP-40 feature, with similar speaker accuracy, thus demonstrating the efficiency of mapping in a low-dimensional latent feature space. In a *habitual* to *clear* style conversion experiment, we showed that VAE-12 together with a custom skip-connection deep neural network significantly improved the speech intelligibility of one of three speakers, with the average keyword recall accuracy increasing from 24% to 46%. In the future, we plan on creating mappings that increase speech intelligibility speaker-independently, using many-to-many voice conversion approaches.

Table 5: *Average keyword accuracy. Results marked with an asterisk are significantly different ($p < 0.05$) as compared to the vocoded HAB condition in a two-tailed t-test.*

	CSM7	PDF7	PDM6
vocoded HAB	38	13	24
DNN-mapping VAE	32	13	35
Skip-mapping VAE	38	11	**46***
hybrid	56*	27*	50*
vocoded CLR	69*	23*	41*

6. References

[1] J. Kates, "Speech enhancement based on a sinusoidal model," *Journal of Speech and Hearing Research*, vol. 37, pp. 449–464, 1994.

[2] Q.-G. Liu, B. Champagne, and P. Kabalab, "A microphone array processing technique for speech enhancement in a reverberant space," *Speech Communication*, vol. 18, pp. 317–334, 1996.

[3] B. Blesser, "Audio dynamic range compression for minimum perceived distortion," *IEEE Transactions on Audio and Electroacoustics*, vol. 17, no. 1, pp. 22–32, 1969.

[4] R. Niederjohn and J. Grotelueschen, "The enhancement of speech intelligibility in high noise levels by high-pass filtering followed by rapid amplitude compression," *IEEE Trans. Audio Speech Lang. Process*, vol. 24, pp. 277–282, 1976.

[5] T. Quatieri and R. McAulay, "Peak-to-rms reduction of speech based on a sinusoidal model," *IEEE Transactions on Audio and Electroacoustics*, vol. 39, pp. 273–288, 1991.

[6] B. Sauert and P. Vary, "Near end listening enhancement: speech intelligibility improvement in noisy environments," pp. 493–496, 2006.

[7] Y. Tang and M. Cooke, "Subjective and objective evaluation of speech intelligibility enhancement under constant energy and duration constraints," *Proceedings of INTERSPEECH*, pp. 345–348, 2011.

[8] T. Takeuchi and Y. Tatekura, "Speech intelligibility enhancement in noisy enviroment via voice conversion with glimpse proportion measure," *Proceeding of APSIPA ASC*, pp. 1713–1717, 2018.

[9] M. Picheny, N. Durlach, and L. Braida, "Speaking clearly for the hard of hearing i: Intelligibility differences between clear and conversational speech," *Journal of Speech and Hearing Research*, vol. 28, pp. 96–103, 1985.

[10] S. H. Ferguson and D. Kewley-Port, "Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners," *Journal of the Acoustical Society of America*, vol. 112, pp. 259–271, 2002.

[11] S. Ferguson, "Talker differences in clear and conversational speech: Vowel intelligibility for normal-hearing listeners," *Journal of the Acoustical Society of America*, vol. 116, pp. 2365–2373, 2004.

[12] A. Kain, A. Amano-Kusumoto, and J.-P. Hosom, "Hybridizing conversational and clear speech to determine the degree of contribution of acoustic features to intelligibility," *Journal of the Acoustical Society of America*, vol. 124, no. 4, pp. 2308–2319, 2008.

[13] K. Tjaden, A. Kain, and J. Lam, "Hybridizing conversational and clear speech to investigate the source of increased intelligibility in speakers with parkinson's disease," *Journal of Speech, language, and hearing research*, vol. 57, pp. 1191–1205, 2014.

[14] S. Mohammadi, A. Kain, and J. van Santen, "Making conversational vowels more clear," *Proceedings of INTERSPEECH*, 2012.

[15] M. Koutsogiannaki, P. N. Petkov, and Y. Stylianou, "Simple and artefact-free spectral modifications for enhancing the intelligibility of casual speech," *Proceedings of ICASSP*, pp. 4648–4652, 2014.

[16] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *Proceedings of ICLR*, 2014.

[17] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from non-parallel corpora using variational auto-encoder," *Proceeding APSIPA ASC*, 2016.

[18] M. Blaauw and J. Bonada, "Modeling and transforming speech using variational autoencoders," *Proceedings of INTERSPEECH*, 2016.

[19] W.-N. Hsu, Y. Zhang, and J. Glass, "Learning latent representations for speech generation and transformation," *Proceedings of INTERSPEECH*, 2017.

[20] M. Morise, F. Yokomori, and K. Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE transactions on information and systems*, vol. E99-D, no. 77, pp. 1877–1884, 2016.

[21] M. Morise, "D4c, a band-aperiodicity estimator for high-quality speech synthesis," *Speech Communication*, vol. 84, no. 57-65, 2016.

[22] L. Deng, X. Cui, R. Pruvenok, J. Huang, S. Momen, Y. Chen, and A. Alwan, "A database of vocal tract resonance trajectories for research in speech processing," *Proceedings of ICASSP*, 2006.

[23] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Proceedings of 3rd ICLR*, 2015.

[24] T. Toda, L.-H. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu, and J. Yamagishi, "The voice conversion challenge 2016," *Proceedings of INTERSPEECH*, 2016.

[25] *International Electrotechnical Commission, Electroacoustics-sound level meters-part 1: Specifications, 61672*, 2002.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," *Proceedings of the ICCV*, pp. 1026–1034, 2015.

[27] K. Tjaden, J. Lam, and G. E. Wilding, "Vowel acoustics in parkinson's disease and multiple sclerosis: comparison of clear, loud, and slow speaking conditions," *Journal of Speech, language, and hearing research*, vol. 56, pp. 1485–1502, 2013.

[28] K. Tjaden, J. E. Sussman, and G. E. Wilding, "Impact of clear, loud, and slow speech on scaled intelligiblity and speech serverity in parkinson's disease and multiple sclerosis," *Journal of Speech, language, and hearing research*, vol. 57, pp. 779–792, 2014.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Computer Vision and Pattern Recognition (CVPR)*, 2016.