INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Whisper Speech Enhancement Using Joint Variational Autoencoder for Improved Speech Recognition

Vikas Agrawal[1], *Shashi Kumar*[1], *Shakti P Rath*[2]

[1]Samsung R&D Institute India – Bangalore, India
[2]Reverie Language Technologies, Bangalore, India

vik.agrawal@samsung.com,sk.kumar@samsung.com and shakti.rath@reverieinc.com

Abstract

Whispering is the natural choice of communication when one wants to interact quietly and privately. Due to vast differences in acoustic characteristics of whisper and natural speech, there is drastic degradation in the performance of whisper speech when decoded by the Automatic Speech Recognition (ASR) system trained on neutral speech. Recently, to handle this mismatched train and test scenario Denoising Autoencoders (DA) are used which gives some improvement. To improve over DA performance we propose another method to map speech from whisper domain to neutral speech domain via Joint Variational Auto-Encoder (JVAE). The proposed method requires time-aligned parallel data which is not available, so we developed an algorithm to convert parallel data to time-aligned parallel data. JVAE jointly learns the characteristics of whisper and neutral speech in a common latent space which significantly improves whisper recognition accuracy and outperforms traditional autoencoder based techniques. We benchmarked our method against two baselines, first being ASR trained on neutral speech and tested on whisper dataset and second being whisper test set mapped using DA and tested on same neutral ASR. We achieved an absolute improvement of 22.31% in Word Error Rate (WER) over the first baseline and an absolute 5.52% improvement over DA.

Index Terms: whisper speech recognition, autoencoder, wTIMIT, Variational autoencoder, jointVAE

1. Introduction

Whisper speech is significantly different from the neutral speech in production. Neutral speech is produced by modulation of airflow from the lungs by vibrations of vocal cords, but there is no vibration of vocal cords in whisper speech. Whisper speech is produced by making a small constriction at the glottis, this results in excitation of the vocal tract without the periodic vibration of vocal cords. Further, because of the lower sound pressure level compared to neutral speech, whisper speech usually has a lower Signal to Noise Ratio (SNR). Also, the frequency of the lower formants of whisper speech is lower than the neutral speech, as well as the spectral tilt of whisper speech is less sloped than neutral speech [1]. All of these makes speech processing, especially recognition of whisper speech more difficult than neutral speech. Generally, speech enhancement techniques like Denoising Autoencoders (DA), Variational Autoencoders (VAE) [2] and Generative Adversarial Networks (GAN) [3] are used for improving speech recognition for whisper data. DA is used to map speech from the whisper domain into the neutral domain using a neural network. GAN is an unsupervised probabilistic generative modelling technique. GANs have been majorly used in the computer vision domain, but now gaining traction in the speech domain too. It has been used for speech synthesis [4], speech enhancement [5] and voice conversion [6] tasks. VAEs model the probabilistic generative process to learn a latent representation [2]. In this paper, we use a variation of VAE for speech enhancement task in the context of whisper ASR. VAE is a class of generative model that projects input space to a latent space using an encoder and then reconstructs the original input using a decoder. VAE is an unsupervised learning technique, in which the output of the network is the same as input and an intermediate latent representation is learned. This latent representation is used to generate more data in the trained domain. In this paper, we propose to explore an extension to VAE, named joint-VAE, as proposed in [7]. Joint-VAE is a supervised learning technique, which uses time aligned parallel data (aligned whisper and neutral speech) to jointly learn the distribution of whisper and neutral speech features using a joint distribution for a common latent space z. In VAE the error term consists of a reconstruction error and a Kullback-Leibler-Divergence (KLD), while in joint-VAE, the error consists of three terms, two reconstruction errors for each whisper and neutral features and KLD. The network consists of an encoder that maps whisper features to latent space z and two decoders mapping z space to whisper features and aligned neutral features. Features generated by decoder corresponding to neutral features are used for decoding in ASR. The rest of the paper is organized as follows. In Section 2, we review related work in the field of whisper speech recognition. In section 3 we describe the conventional autoencoder. In section 4 we present our proposed architecture of joint-VAE. In section 5 we describe our experimental setup and results on the wTIMIT dataset. Section 6 discusses the conclusion.

2. Related Work

Automated whisper speech recognition task is a difficult task and is also marred by lack of availability of data. In general, there are two strategies to handle whisper speech.

The first strategy is to adapt the model trained on neutral speech to whisper domain, to do this in [8] a model to capture whisper speech characteristics from a small amount of whisper data and then used the Vector Taylor Series (VTS) algorithm to transform neutral speech to whisper speech. These whisper speech data generated was used to adapt the model to the whisper domain. In [9], VTS along with GAN was used to generate pseudo-whisper data to adapt the model to the whisper domain. In [1], the model is trained with whisper data and Maximum Likelihood Linear Regression (MLLR) adaptation is applied to increase the accuracy.

http://dx.doi.org/10.21437/Interspeech.2021-953

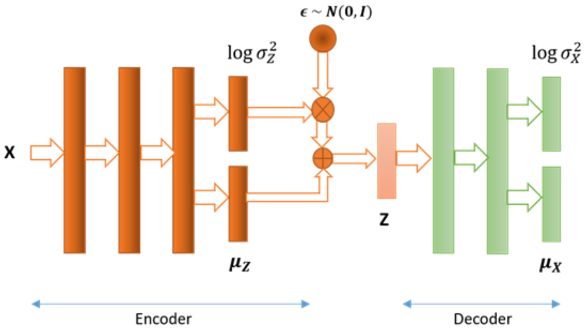

Figure 1: *Conventional VAE architecture.*

The second strategy is to map the whisper features to neutral speech features and decoding the generated features with a model trained on neutral speech. To do this in [10], trained a neural network that redistributed filterbanks for features adapted to whisper and tested those on Hidden Markov Model (HMM) based system. In [11], features were transformed using various techniques like MLLR, feature space adaptation, retraining with downsampling, sigmoid low pass filter and linear multivariate regression. In [12], four Bi-directional Long-Short Term Memory (BLSTM) networks were used for feature, pitch, periodicity and Voiced/Unvoiced phoneme decision, for mapping from whisper speech to neutral speech. In [13], inverse filtering technique was used to enhance whisper speech. In [14], radial basis function neural network (RBF NN) was used to model whisper speech to neutral speech. In [15], discoGAN a variation of GAN was used to map whisper speech to neutral speech. In [16], denoising autoencoder is used to map whisper features to neutral speech features.

3. Review of Variational Autoencoder

In recent years, VAEs have gained traction for unsupervised learning of unknown and intractable distributions. These models are essentially encoder-decoder based models where the encoder maps input feature space to latent space and the decoder tries to reconstruct the features given samples from latent space. Standard VAE tries to reconstruct the input space and hence it does not offer domain translation.

Figure 1 shows a conventional VAE architecture, the encoder encodes the inputs to a latent space and the decoder decodes these latent encodings to the original input.

Let the input observed data be $X : x_1, x_2 ... x_n$. VAE assumes the observed data is generated by some latent space variable $Z : z_1, z_2 ... z_n$. In order to learn the generative process, we measure the posterior distribution $p_\theta(z|x)$ where p_θ denotes the probability distribution parameterized by θ. From Baye's rule

$$p_\theta(z|x) = \frac{p_\theta(x|z)p_\theta(z)}{p_\theta(x)} \tag{1}$$

The posterior probability $p_\theta(z|x)$ in (1) is intractable for most of the distributions, as it involves computation of multi-dimensional integral:

$$p_\theta(x) = \int p_\theta(x|z)p_\theta(z)dz \tag{2}$$

A better solution is to approximate $p_\theta(z|x)$ by another distribution $q_\phi(z|x)$ which is tractable. To make $q_\phi(z|x)$ as close as possible to $p_\theta(z|x)$, we try to minimize the KLD between the two, i.e.,

$$minKL(q_\phi(z|x)||p_\theta(z|x)) \tag{3}$$

Then, it is easy to show that

$$logp_\theta(x) = \mathcal{L}_1(\theta, \phi; x) + KL(q_\phi(z|x)||p_\theta(z|x)) \tag{4}$$

$$logp_\theta(x) \geq \mathcal{L}_1(\theta, \phi; x) \tag{5}$$

where $p_\theta(x)$ denotes the marginal distribution of the observed data and $\mathcal{L}_1(\theta, \phi; x)$ denotes the variational lower bound, which is defined as

$$\mathcal{L}_1(\theta, \phi; x) = \int_z q_\phi(z|x)log\frac{p_\theta(x,z)}{q_\phi(z|x)} \tag{6}$$

$$\mathcal{L}_1(\theta, \phi; x) = E_{q_\phi(z|x)}logp_\theta(x|z) - KL(q_\phi(z|x)||p_\theta(z)) \tag{7}$$

The parameters θ and ϕ are jointly optimized. The negative of the variational lower bound of the marginal distribution is minimized. Both distribution $p_\theta(z|x)$ and $q_\phi(z|x)$ are modelled by diagonal Gaussian distributions given by

$$p_\theta(z|x) = \mathcal{N}(x; f_{\mu_z}(z;\theta), exp(f_{log\sigma_z^2}(z;\theta))) \tag{8}$$

$$q_\phi(z|x) = \mathcal{N}(z; g_{\mu_x}(z;\phi), exp(g_{log\sigma_x^2}(x;\phi))) \tag{9}$$

Generally, neural networks are used to model mean and log variance of (8) and (9). Prior $p_\theta(z)$ is modelled by isotropic multi-variate Gaussian distribution

$$p_\theta(z) = \mathcal{N}(z; 0, I) \tag{10}$$

To compute expectation in (7), we draw a sample $\hat{z}$ from $q_\phi(z|x)$ and setting $E_{q_\phi(z|x)}logp_\theta(x|z) \approx logp_\theta(x|\hat{z})$. Due to sampling, the network becomes dis-continuous and thus non-differentiable and hence backpropagation can not be applied for training. To overcome this, we use re-parameterization trick [2], latent variable z is given as

$$z = f_{\mu_z}(z;\theta) + \sqrt[2]{exp(f_{log\sigma_z^2}(z;\theta))} \odot \epsilon \tag{11}$$

where $\odot$ denotes element-wise multiplication and ϵ is sample drawn from $\mathcal{N}(z; 0, I)$.

We note here that similar to DA, in practice, variational autoencoder possibly extended for domain translation, where normal speech may be reconstructed given whisper speech as input. Unfortunately, such operation cannot be supported theoretically as the input and output to VAE must be the same, as evident from (7). Using a different input and output breaks the VAE framework. However, the authors in [17] have explored VAE without considering the above fact. We believe the lack of a theoretical formulation that supports VAE may lead to unpredicted results. Therefore, the results are not shown for VAE in this work.

4. Proposed joint VAE

Figure 2 shows our proposed joint-VAE architecture, we have an encoder and two decoders jointly trained for neutral and whisper domain. We assumed the availability of time-aligned paired data. We denote whisper speech data by x and neutral speech data by y. z being the latent space. We modify the variational lower bound as defined in (7) to incorporate both decoders as

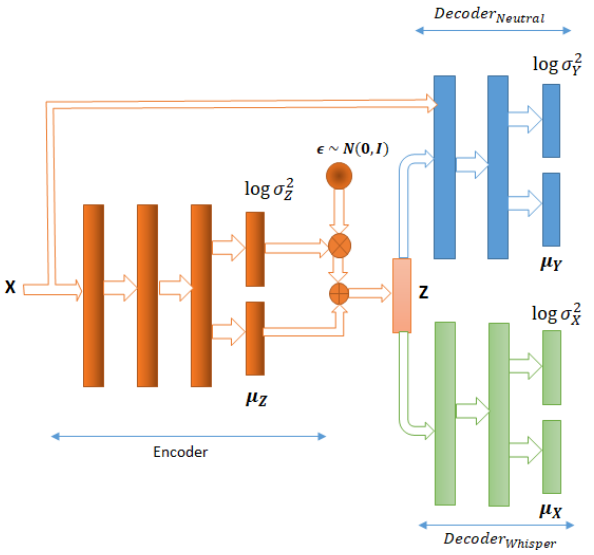

Figure 2: *Proposed Joint-VAE architecture.*

$$\mathcal{L}_2(\theta, \phi; x, y) = \int_z q_\phi(z|x, y) log \frac{p_\theta(x, y, z)}{q_\phi(z|x, y)} \tag{12}$$

$$\begin{aligned} \mathcal{L}_2(\theta, \phi; x, y) = E_{q_\phi(z|x)} log p_\theta(x|z) \\ + E_{q_\phi(z|x)} log p_\theta(y|x, z) - KL(q_\phi(z|x)||p_\theta(z)) \end{aligned} \tag{13}$$

We made an approximation of $q_\phi(z|x, y) = q_\phi(z|x)$ assuming the mapping between whisper speech features and neutral speech features is deterministic. Now, we have two conditional distributions, $p_\theta(y|x, z)$ and $p_\theta(x|z)$, and one posterior distribution, $q_\phi(z|x)$, each modelled by a diagonal gaussian distribution.

$$p_\theta(x|z) = \mathcal{N}(x; f^1_{\mu_z}(z; \theta), exp(f^1_{log\sigma_z^2}(z; \theta))) \tag{14}$$

$$p_\theta(y|x, z) = \mathcal{N}(y; f^2_{\mu_{x,z}}(x, z; \theta), exp(f^2_{log\sigma^2_{x,z}}(x, z; \theta))) \tag{15}$$

$$q_\phi(z|x) = \mathcal{N}(z; g_{\mu_x}(x; \phi), exp(g_{log\sigma_x^2}(x; \phi))) \tag{16}$$

As in VAE, the mean and log variance are modelled by neural networks and the prior $p_\phi(z)$ is modelled by isotropic gaussian. In practice, the loss optimized is given below

$$\mathcal{L}_3 = \lambda_1 MSE_{whisper} + \lambda_2 MSE_{neutral} + \lambda_3 KLD \tag{17}$$

In (17), the first term $MSE_{whisper}$ is defined as the mean square error between reconstructed whisper output and whisper input, the second term $MSE_{neutral}$ is the mean square error between reconstructed neutral speech features and neutral output features. The third term denotes KL-divergence between $q_\phi(z|x)$ and prior distribution $p_\theta(z)$. A weighted sum of all three error terms is taken to arrive at total error and the weights λ_1, λ_2 and λ_3 are tuned experimentally using grid search. The KLD term forces the distribution $q_\phi(z|x)$ to be as close as distribution $p_\theta(z)$ leading to smoothed distribution in latent space. The reconstruction terms encourage the distributions in the latent space to be as granular as possible while enabling the latent variables to deviate from prior distribution when necessary.

5. Experiments and Results

We used wTIMIT [18] whisper speech corpus for our experiments. The corpus is in English and was collected in two phases. All speech utterances were recorded in a clean acoustic environment at 44.1 kHz. There are a total of 48 speakers in the corpus. The corpus is divided into train and test sets and each of these sets contains both neutral and whisper speech data. We have about 20 hrs of clean and 20 hrs of whisper speech data in the train set. The test set contains 1263 whisper and 1265 neutral utterances.

We used the standard Kaldi toolkit [19] recipe for Wall Street Journal (WSJ) task to train the GMM-HMM model on whisper and neutral data combined. We train with both whisper and neutral data to generate alignments for both the sets. The features are 13-dimensional Mel Frequency Cepstral Coefficients (MFCC). Cepstral mean normalization is applied to the MFCC features on a per-speaker basis. Afterwards, splicing is applied taking a context of ±4 and linear discriminant analysis is applied for dimensionality reduction and de-correlation of the features. The resulting features were further diagonalized using Semi-Tied Co-variance (STC) [20] (also known as maximum likelihood linear transform (MLLT) [21]. Then Speaker Adaptive Training (SAT) was applied using Constrained Maximum Likelihood Linear Regression (CMLLR) [22]. The GMM-HMM is trained in the LDA-MLLT-SAT space.

5.1. Generation of Time Aligned Parallel Data

In wTIMIT, a speaker has spoken a sentence both in a neutral voice and whisper voice but these utterances are not time-aligned. By time aligned we mean that for each frame x_n of whisper we know the time-synchronous frame y_n of neutral speech. To align them first, we generated senone alignments for both whisper and neutral speech as explained in previous section. Next, we converted senone alignments to phone alignments. Now for each utterance first we checked for the sequence of phones in both whisper and neutral alignments. We picked the first phone from both sequences and check if they are the same or not. If not the same then either of them is a silence phone, we drop all silence phones and pick the next phone from that sequence. If same phone then we checked the number of frames corresponding to that phone in both alignments. If the number of frames are the same then we checked the next phone in sequence and continue, else we checked which alignment has extra frames, we arranged all frames corresponding to that phone from that alignment in increasing order of energy. Then we removed extra frames which have the least energy while maintaining the order of frames in alignment.

5.2. ASR model

We trained the GMM-HMM LDA+MLLT+SAT system as above but only using neutral speech data and generated alignments for acoustic model (AM) training. The AM is a long short term memory (LSTM) network trained on neutral speech data. We use 41-dimensional log-mel filterbank features with splicing of ±2 as an input for the experiments. For better cell state initialization, the left context of 20 frames is used. The network consists of 3 layers of LSTM with 512 hidden cells using context-dependent states obtained from the GMM-HMM model as targets. Adam optimizer is used for training with a learning rate varying from 0.001 to 0.0001 exponentially for the first 20 epochs. The learning rate is kept fixed for the remaining 5 epochs. We fix the LSTM-HMM model trained on neu-

Table 1: *WER (%) and SER (%) on wTIMIT whisper test set and neutral test set*

Train	Test	WER(%)	SER(%)
Neutral	Neutral	0.99	5.93
Neutral	Whisper	31.17	63.90

Table 2: *WER (%) and SER(%) on wTIMIT whisper test set for different DA configurations*

DA Layers	Train	Test	WER(%)	SER(%)
3	Neutral	Whisper	14.95	39.17
5	**Neutral**	**Whisper**	**14.38**	**38.88**
7	Neutral	Whisper	16.03	41.41

tral speech data as our standard recognition system and whisper test set as a standard test scenario. The word error rate (WER) is 31.17% and the sentence error rate (SER) is 63.90% on the wTIMIT whisper test set as shown in Table 1. For the wTIMIT neutral test set, WER is 0.99% and SER is 5.93%.

5.3. Denoising Autoencoder

Our second baseline is DA which maps whisper features to neutral features. We experimented with 3 layers, 5 layers and 7layer LSTM followed by one fully connected layer. Our best results are from 5layer LSTM as shown in Table 2, hereafter this result is referred to as DA result. The mean square error between neutral features and the predicted neutral speech features is minimized to train this DA. We used stochastic gradient descent (SGD) with a momentum of 0.9 to minimize this error. During testing, the test set of whisper features are passed through this DA and the enhanced output is decoded using the LSTM-HMM ASR model. The results are shown in Table 3, from which we note that the WER of this model on the whisper test set was 14.38%.

5.4. Proposed joint-VAE

The encoder network of joint-VAE comprises 3 layers LSTM each consisting of 512 hidden cells, followed by two parallel fully connected layers, one each for mean and log-variance. The activation in these output layers are linear units as the range mean and log-variance are unbounded. The encoder takes whisper data as input and predicts mean and log variance of latent space z. $Decoder_{whisper}$ consists of 2 LSTM layers followed by two parallel fully connected layers with linear activations, which takes latent vector z as input and outputs mean and log-variance of reconstructed whisper data. The architecture of $Decoder_{neutral}$ is similar, which takes both latent vector and original whisper data as inputs and predicts mean and log-variance of reconstructed neutral data. Batch normalization is applied to every layer except fully connected layers. The complexities of encoder and decoder networks are kept different to avoid the possibility of learning identity or orthogonal mapping.

In our proposed architecture for joint-VAE, the loss function as defined in (17) is minimized using stochastic gradient descent (SGD) with a momentum of 0.9. The learning rate is fixed at 0.001 for the first 30 epochs and 0.0001 for the remaining 20 epochs. Hyperparameters λ_1, λ_2 and λ_3 were fine-tuned experimentally. During test time, we decode the mean predicted by the $Decoder_{neutral}$ of the joint-VAE network using the acoustic model discussed above. We achieved WER of 8.86% on the wTimit whisper test set as shown in Table 3 which is an absolute improvement of 5.52% over the baseline DA model and at the same time SER is improved by 11.48%.

Table 3: *WER (%) and SER(%)on wTIMIT whisper test set on DA and jVAE*

Train	Model	WER(%)	SER(%)
Neutral	DA	14.38	38.88
Neutral	**jVAE**	**8.86**	**27.40**

Table 4: *WER (%) and SER(%) on wTIMIT whisper test set on jVAE with different hyperparameters*

λ_1	λ_2	λ_3	WER(%)	SER(%)
1	1	1	14.41	37.69
2	20	2	13.16	35.47
2	20	0.2	11.92	34.76
2	**20**	**0.1**	**8.86**	**27.40**

5.5. Finetuning of hyperparameters

Tuning the hyperparameters λ_1, λ_2 and λ_3 defined in (17) is important, Table 4 shows the results when different weights were given to MSE and KL terms. With the equal value of parameters, we get WER of 14.41% which is slightly better than DA. Since our focus is on the reconstruction of neutral speech, we boosted λ_2 to 20 which led to significant improvement in WER. Following [7], we lowered the value of λ_3 to 0.2 to encourage clustering over smoothing between classes. This improved the performance significantly to 11.92%. To achieve parity between the absolute value of errors in (17), we lowered the value of λ_3 further to 0.1. As evident from the fourth row of Table 4, we achieve the best performance using this combination of hyperparameters. Specifically, we observed a relative improvement of 38% in WER from DA.

6. Conclusion

This paper presented an extension of the conventional variational autoencoder, named joint-VAE, that enables joint distribution learning of observations from two different sources. The method is evaluated for speech enhancement in the context of whisper speech recognition, where the two observations are a sequence of whisper and neutral features aligned in time. In contrast to conventional VAE, mathematical developments shown in this paper depicts that the variational lower bound associated with joint-VAE involves two reconstruction errors, one corresponding to whisper features and another to neutral features. For speech enhancement, the whisper features are passed through the encoder network to generate the latent variables, which are later decoded by the two decoder networks to reconstruct whisper and neutral features. Experiments conducted on the wTIMIT dataset shows that joint-VAE yields an absolute WER improvement of 5.52% and absolute SER improvement of 11.48% compared to the conventional DA and absolute WER improvement of 22.31% compared to the baseline whisper test set.

7. References

[1] T. Itoh, K. Takeda, and F. Itakura, "Acoustic analysis and recognition of whispered speech," in *IEEE Workshop on Automatic Speech Recognition and Understanding, 2001. ASRU '01.*, Dec 2001, pp. 429–432.

[2] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[3] 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*, 2014, pp. 2672–2680.

[4] S. Pascual, A. Bonafonte, and J. Serra, "Segan: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[5] 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," *arXiv preprint arXiv:1704.00849*, 2017.

[6] T. Kaneko, H. Kameoka, N. Hojo, Y. Ijima, K. Hiramatsu, and K. Kashino, "Generative adversarial network-based postfilter for statistical parametric speech synthesis," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 4910–4914.

[7] M. K. Chelimilla, S. Kumar, and S. P. Rath, "Joint distribution learning in the framework of variational autoencoders for far-field speech enhancement," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 245–251.

[8] S. Ghaffarzadegan, H. Boril, and J. H. Hansen, "Model and feature based compensation for whispered speech recognition," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[9] S. Ghaffarzadegan, H. Bořil, and J. H. Hansen, "Generative modeling of pseudo-whisper for robust whispered speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 10, pp. 1705–1720, 2016.

[10] S. Ghaffarzadgan, H. Bořil, and J. H. Hansen, "Ut-vocal effort ii: Analysis and constrained-lexicon recognition of whispered speech," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 2544–2548.

[11] S.-C. Jou, T. Schultz, and A. Waibel, "Whispery speech recognition using adapted articulatory features," in *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 1. IEEE, 2005, pp. I–1009.

[12] G. N. Meenakshi and P. K. Ghosh, "Whispered speech to neutral speech conversion using bidirectional lstms." in *Interspeech*, 2018, pp. 491–495.

[13] D. T. Grozdić, S. T. Jovičić, J. Galić, and B. Marković, "Application of inverse filtering in enhancement of whisper recognition," in *12th Symposium on Neural Network Applications in Electrical Engineering (NEUREL)*, Nov 2014, pp. 157–162.

[14] Z. Tao, X.-D. Tan, T. Han, J.-H. Gu, Y.-S. Xu, and H.-M. Zhao, "Reconstruction of normal speech from whispered speech based on rbf neural network," in *2010 Third International Symposium on Intelligent Information Technology and Security Informatics*. IEEE, 2010, pp. 374–377.

[15] M. Parmar, S. Doshi, N. J. Shah, M. Patel, and H. A. Patil, "Effectiveness of cross-domain architectures for whisper-to-normal speech conversion," in *2019 27th European Signal Processing Conference (EUSIPCO)*. IEEE, 2019, pp. 1–5.

[16] T. Grozdic and T. Jovicic Slobodan, "Whispered speech recognition using deep denoising autoencoder," *Engineering Applications of Artificial Intelligence*, vol. 59, pp. 15–22, 2017.

[17] Y. Jung, Y. Kim, Y. Choi, and H. Kim, "Joint learning using denoising variational autoencoders for voice activity detection." in *INTERSPEECH*, 2018, pp. 1210–1214.

[18] P. Bauer, D. Scheler, and T. Fingscheidt, "Wtimit: The timit speech corpus transmitted over the 3g amr wideband mobile network." in *LREC*, 2010.

[19] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, dec 2011, iEEE Catalog No.: CFP11SRW-USB.

[20] M. J. F. Gales, "Semi-tied covariance matrices for hidden markov models," *IEEE Trans. Speech and Audio Processing*, vol. 7, pp. 272–281, 1999.

[21] R. A. Gopinath, "Maximum likelihood modeling with gaussian distributions for classification," in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)*, vol. 2, 1998, pp. 661–664 vol.2.

[22] M. J. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," *Computer speech & language*, vol. 12, no. 2, pp. 75–98, 1998.