Perceptual Characteristics Based Multi-objective Model for Speech Enhancement

Chiang-Jen Peng[1], *Yih-Liang Shen*[1], *Yun-Ju Chan*[1], *Cheng Yu*[2], *Yu Tsao*[2], *Tai-Shih Chi*[1]

[1]Department of Electrical and Computer Engineering, National Yang Ming Chiao Tung University
Hsinchu, Taiwan
[2]Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan

yu.tsao@citi.sinica.edu.tw, tschi@nycu.edu.tw

Abstract

Deep learning has been widely adopted for speech applications. Many studies have shown that using the multiple objective framework and learned deep features is effective for improving system performance. In this paper, we propose a perceptual characteristics based multi-objective speech enhancement (SE) algorithm that combines the conventional loss and objective losses of pitch and timbre related features. Timbre related features include frequency modulation (encoded by the pitch contour), amplitude modulation (encoded by the energy contour), and speaker identity. For the speaker identity loss, we consider the deep features derived in a speaker identification system. The proposed algorithm consists of two parts, a LSTM based SE model and CNN based multi-objective models. The objective losses are derived between speech enhanced by the SE model and clean speech and combined with the SE loss for updating the SE model. The proposed algorithm is evaluated using the corpus of Taiwan Mandarin hearing in noise test (TMHINT). Experimental results show the proposed algorithm evidently outperforms the original SE model in all objective scores, including speech quality, speech intelligibility and signal distortion.

Index Terms: Speech enhancement, perceptual characteristics, timbre, deep feature, multi-objective model

1. Introduction

In daily life, environmental noise causes serious degradation to speech quality and intelligibility, thus reduces the performance of speech application systems, such as assistive hearing systems [1][2], speaker recognition systems [3][4], and automatic speech recognition systems [5][6]. Numerous speech enhancement (SE) or noise reduction methods have been proposed through decades. Conventional SE techniques, such as subspace decomposition [7], power spectral subtraction [8], Wiener filtering [9], and minimum mean square error (MMSE) based estimations [10], perform well in stationary noise but unsatisfactorily in non-stationary noise, which commonly exists in daily environments. During the past decade, neural network (NN) models have been shown capable of dramatically improving performance of SE systems in non-stationary noise. Some early studies demonstrated deep-learning (DL) based NN models provide significant improvement over conventional methods even only working on the magnitude of the spectrogram [11][12][13]. Recently, the DCCRN model [14], which combines the complex-value CNN model and LSTM, and the two-stage model [15] were proposed to deal with the magnitude and phase of the spectrogram simultaneously. On the other hand, some methods were proposed to directly enhance speech in the time domain to bypass the difficulty of handling the phase [16][17] .

Not only SE, voice conversion (VC) also has the goal to produce high quality speech signals. VC is a task of synthesizing speech with converted speaker characteristics from a source speaker to a target speaker and preserved linguistic contents of source speech [18]. To ensure the converted speech sounds like from the target speaker, methods have been proposed to disassemble the speech signal into many different characteristics and use different methods to convert these characteristics one by one [19][20]. Intuitively, the more detailed characteristics the sound disassembled into, the better chance of obtaining a converted sound with high quality. Similarly, researchers in the SE field began to consider certain sound characteristics in their SE methods since quality of enhanced speech is critical for SE systems. For instance, the speaker embedding trained by a speaker identification model was added into the SE model to improve model performance [21][22]. Other characteristics like the fundamental frequency (F0), energy or even contextual information were considered to improve performance of SE methods [23][24][25].

Inspired by these approaches, we propose a perceptual characteristics based multi-objective SE model in this paper. Loudness, pitch and timbre are the three prominent perceptual characteristics of the sound. Comparing with loudness and pitch, timbre is much more complicated. It contains spectral, temporal and spectral-temporal elements. Estimating spectral envelope per frame is the usual approach in conventional SE methods. The spectral envelope per frame purely conveys spectral information of timbre. As for timbre's temporal elements, we consider the amplitude modulation and the frequency modulation of speech as important elements. In addition, we consider the spectral-temporal content for identifying the speaker is an important spectral-temporal element of timbre. Therefore, we investigate four features, pitch (fundamental frequency), the pitch contour (encoding the frequency modulation), the energy contour (encoding loudness and the amplitude modulation), and the spectral-temporal feature for speaker identity, in the proposed multi-objective SE model. These features are first extracted by feature models separately. Then, a composite loss function, which contains the conventional spectral-envelope loss and objective losses of these features, is used to update the multi-objective SE model. The proposed model is evaluated using the Taiwan Mandarin hearing in noise test (TMHINT) corpus [26]. We use Mandarin dataset since frequency modulation is more prominent in tonal languages than in non-tonal languages. Experimental results clearly show the proposed multi-objective

This research is partly supported by the Ministry of Science and Technology, Taiwan under Grant No MOST 107-2221-E-009-132-MY3.

model outperforms the plain SE model in tested scores, including speech quality, speech intelligibility and signal distortion.

The remainder of this paper is organized as follows. Section 2 introduces the perceptual characteristics based features and the feature models. Section 3 introduces the proposed multi-objective SE method. Experimental results and analyses are provided in Section 4. Finally, Section 5 concludes the paper and gives directions for future work.

2. Perceptual characteristics based features

This section introduces the perceptual characteristics based features used in our method and their corresponding feature models.

2.1. Pitch (fundamental frequency) feature

Based on the definition given by American Standards Association (ASA), pitch is the attribute of auditory sensation in terms of which sounds may be ordered on a musical scale. It is related to the repetition rate of the waveform. For complex sounds as speech, pitch usually corresponds to the fundamental frequency (F0). Without considering the phenomenon of residue pitch nor unvoiced pitch of whispered speech in this paper, the terms of pitch and F0 are interchangeable.

2.2. Timbre related features: Energy contour, pitch contour, and speaker identity

According to ASA, timbre is the attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar. In other words, timbre includes everything else besides loudness and pitch. Therefore, timbre contains spectral, temporal and spectro-temporal elements. Since the spectral envelope per frame has already been used as the loss function in our baseline SE system, we aim to add temporal and spectro-temporal elements of timbre as additional objectives to improve system performance.

For the temporal elements of timbre, we consider amplitude modulation and frequency modulation, which are well known important temporal cues in many tasks, such as auditory streaming [27][28]. In this paper, we adopt the energy contour and the pitch contour as temporal features for encoding amplitude modulation and frequency modulation, respectively. These temporal features contain lot of information of the speech signal, such as intonation, melody, rhythm and its emotion state. Not only that, they also provide long-term information to our SE method, which is complementary to the short-term information provided by the frame-wise spectral envelope.

In addition, from daily life experience, speaker identity provides abstract information of timbre to help people distinguish utterances easily. In this study, we consider speaker identity is embedded in the spectro-temporal structure of speech, hence, it can be thought as a spectro-temporal element of timbre.

2.3. Feature models

We use two 4-layer 1-D CNN models to estimate F0 and energy per frame. They are then collected across frames to form the estimated pitch and energy contours. The ground truth of F0 and energy were derived using the auto-correlation function and the L2-norm in each speech frame. To capture spectro-temporal information of timbre embedded in speaker identity, we build a speaker identification model using a 4-layer 2-D CNN model. The accuracy of our speaker identification model on TMHINT corpus (8 speakers) is 100 percent. The output of the layer before the final output of the speaker identification model is used as the deep feature to represent the speaker identity. Details of these feature models are summarized in Table 1.

Table 1: *Architecture of models. FC[hidden units] is short for fully-connected layer, and $Conv$[kernel size, kernel numbers, stride] is short for convolution layer.*

Model	Baseline SE	Speaker ID	Pitch/Energy
Input size	257×N	257×N	257
Layer1	$LSTM[300]$	$Conv[1 \times 5, 256, 1]$	$Conv[3, 256, 1]$
Layer2	$LSTM[300]$	$Conv[1 \times 7, 256, 2]$	$Conv[3, 256, 1]$
Layer3	$FC[257]$	$Conv[1 \times 1, 256, 1]$	$Conv[3, 256, 1]$
Layer4		$Conv[1 \times 1, 512, 1]$	$FC[1]$
Layer5		$FC[256]$	
Layer6		$FC[8]$	
Output size	257×N	8	1

3. The proposed model

In this section, we describe details of the baseline SE model and the proposed multi-objective SE model.

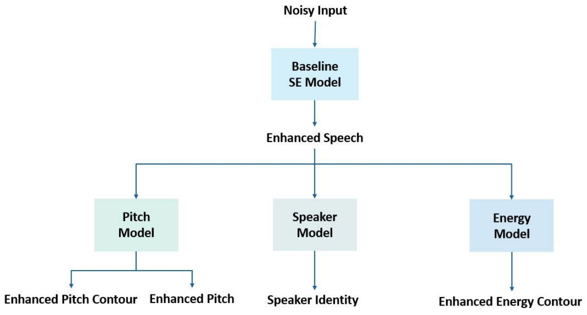

Figure 1: *The block diagram of the proposed SE model. The input is noisy speech while the outputs are enhanced speech and the enhanced perceptual characteristics based features.*

3.1. Baseline SE model

The baseline spectral-feature based SE model is constructed with two LSTM layers followed by one linear layer. The LSTM layers are trained sequence by sequence with the hidden size of 300. The output dimension is set to the hidden size, while the output time step remains the same as the input. The following linear layer transforms the output of the second LSTM layer into the spectral feature of enhanced speech, $\tilde{X}$. Details of the baseline SE model is also shown in Table 1.

To train this spectral-feature based baseline model, we first prepared a training set which is composed of noisy-clean pairs of speech spectra. Secondly, we trained the model via supervised training by minimizing the MSE loss between the enhanced spectral feature $\tilde{X}$ and the clean spectral feature while the noisy spectral feature was used as the input.

3.2. The proposed multi-objective SE model

Figure 1 shows the block diagram of the proposed multi-objective SE model. The input to the SE model is a noisy logarithmic power spectrum (LPS) feature, and the output is the enhanced LPS feature. The enhanced LPS feature is then fed into three feature models to produce the enhanced perceptual characteristics based features.

In SE literature, many studies only minimize MSE between the enhanced LPS and the clean LPS with the following loss function:

$$Loss_{LPS} = \frac{1}{N^{(I)}} \sum_{t=0}^{N^{(I)}-1} \left(\frac{1}{F} \left\| X_t^{(I)} - \tilde{X}_t^{(I)} \right\|_2 \right) \tag{1}$$

where X is the clean LPS and $\tilde{X}$ is the enhanced LPS. I and t represent the training pair index and the frame index. N and F are the total numbers of the time frame and the frequency bin of the LPS. In our opinion, using this loss function to update the SE model is not good enough. The better way should be preserving perceptual characteristics of speech when minimizing the loss between the enhanced and the clean LPS. Thus, we not only keep the conventional LPS loss but also add objective losses from those mentioned perceptual characteristics based features. For objective losses of pitch and speaker identity, we use the MSE criteria to calculate the loss between clean and enhanced speech. For objective losses of the pitch contour and the energy contour, we use the cosine similarity measure to observe the difference of the long-term features between clean and enhanced speech. Finally, the composite loss function is formulated as follows:

$$\begin{aligned} Loss = & Loss_{LPS} + \frac{1}{N^{(I)}} \sum_{t=0}^{N^{(I)}-1} \left(\left\| P_t^{(I)} - \tilde{P}_t^{(I)} \right\|_2 \right) \\ & + \frac{1}{M} \left\| S^{(I)} - \tilde{S}^{(I)} \right\|_2 + cos(E^{(I)}, \tilde{E}^{(I)}) \\ & + cos(R^{(I)}, \tilde{R}^{(I)}) \end{aligned} \tag{2}$$

where P is the ground truth pitch, S is the M-dimensional deep features of speaker identity derived from clean speech, E is the energy contour of clean speech, R is the pitch contours of clean speech, and cos represents the cosine similarity measure. $\tilde{P}$, $\tilde{S}$, $\tilde{E}$, and $\tilde{R}$ are the estimated counterparts of P, S, E, and R from enhanced speech by the proposed feature models.

4. Experimental results

4.1. Experimental setup

4.1.1. Dataset

We evaluated the proposed SE model using TMHINT utterances, recorded in a noise-free meeting room with 16 kHz sampling frequency. Utterances of six out of eight speakers were used to build the training, validation and test sets. A total of 880 clean utterances spoken by three male and one female speakers were selected for the training purpose. Noisy utterances were generated by adding 24 types of noise, each of which was added at a random signal-to-noise ratio (SNR) between -10 dB and 20 dB. In the end, we generated 21120 ($= 880 \times 24$) noisy-clean utterance pairs for training. Among them, we randomly selected 2112 noisy-clean pairs to form the validation set. In the test set, different 200 clean utterances spoken by one male and one female were mixed with three types of additive noise (engine, white, and street noise) at 10 SNRs (-7, -6, -3, -1, 0, 1, 3, 4, 6 and 9 dB). These three types of noise are unseen to the training and validation sets. Accordingly, 6000 ($= 200 \times 3 \times 10$) noisy utterances were prepared for test.

4.1.2. Experiment settings and evaluation metrics

We applied the short-time Fourier transform (STFT) to speech signals with the window size of 32 ms and the 16 ms frame shift to obtain 257-dimensional LPS vectors. In training, the batch size was set to 128, and the learning rate was set to 0.005 for the baseline SE model. The perceptual characteristics based feature models were also finetuned by the composite loss function so that all involved models would collectively produce the optimal solution. The learning rate for finetuning was set to 0.001. Early stop was triggered when no further loss decreasing on the validation set for 10 epochs. Adam optimizer [29] was used and CosineAnnealingLR was used as the learning rate scheduler. Besides, to prevent the noise-like output of the baseline SE model during early training from destroying the pre-trained perceptual characteristics based feature models, we only used the LPS loss as in Eq. (2) to train the baseline SE model during the first 5 training epochs. After 5 training epochs, the perceptual characteristics based objective losses were added into the loss function as in Eq. (3) for training the multi-objective SE model.

Table 2: *Averaged PESQ, ESTOI, CSIG and COVL scores of noisy utterances, and enhanced utterances by the baseline model and the proposed multi-objective SE model.*

	PESQ	**ESTOI**	**CSIG**	**COVL**
Noisy	1.4	0.48	1.48	1.25
Baseline model	1.72	0.47	2.17	1.58
Multi-objective model	1.83	0.49	2.28	1.68

We adopted four metrics for performance evaluation, perceptual evaluation of speech quality (PESQ) [30], extended short time objective intelligibility (ESTOI) [31], composite signal distortion (CSIG), and composite overall quality using the scale of the mean opinion score (COVL) [32]. Higher PESQ and ESTOI scores indicate better speech quality and intelligibility. Higher CSIG and COVL scores indicate better signal-level SE performance and overall speech quality.

4.2. Experimental results

Table 2 presents the averaged PESQ, ESTOI, CSIG and COVL scores over all noisy utterances, and all enhanced utterances by the baseline SE model and the proposed multi-objective model. From the table, the baseline SE model produces enhanced utterances with better PESQ, CSIG and COVL scores than original noisy utterances. The results demonstrate the effectiveness of the baseline model in improving speech quality and reducing background noise. Meanwhile, the proposed multi-objective SE model provides the best results in all evaluation metrics. The results confirm that the baseline SE model can be further improved by adding perceptual characteristics based feature losses. Figure 2 shows magnitude spectrograms of a clean sample utterance, of the enhanced utterances by the baseline model and the proposed multi-objective model from top to bottom panels, respectively. The bottom two panels show the proposed model preserve more speech components and produce less residual noise in silent parts than the baseline model.

For further analysis, we evaluated the proposed model with each of the perceptual features. Experimental results are shown in Table 3 and Table 4. The bottom part of Table 3 provides

Table 3: *Averaged evaluation scores produced by the baseline model with each proposed perceptual feature.*

	PESQ	ESTOI	CSIG	COVL
Noisy	1.4	0.48	1.48	1.25
Baseline model (BM)	1.72	0.47	2.17	1.58
BM + Speaker Identity	1.82	0.49	2.2	1.62
BM + Pitch	1.83	0.49	2.23	1.66
BM + Pitch Contour	1.87	0.49	2.25	1.67
BM + Energy Contour	1.89	0.5	2.25	1.67

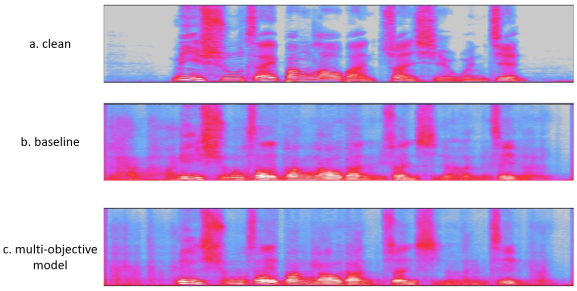

Figure 2: *Spectrograms of a clean utterance, enhanced utterances by the baseline model and the proposed model.*

scores of the baseline model combined with each perceptual feature model. Table 4 demonstrates the average improvement (in percentage) over all evaluation scores when adding each perceptual feature to the baseline model. Clearly, adding any perceptual feature can improve the performance of the baseline model. Results also show that the energy contour added model produces the highest scores in all of the evaluation metrics. In addition, one can observe that the top two performance boosters are the energy contour and the pitch contour, which contain long-term timbre information. Although the 4-layer 2-D CNN speaker identity model can extract spectro-temporal cue of timbre, the word-level spectro-temporal information is still quite local comparing with the sentence-level contours.

In addition to investigate the performance gain from each perceptual feature, we also examined the gains from all possible combinations of two features and three features. The bottom part of Table 5 shows the best performance of adding one feature, of adding two features, of adding three features, and the performance of adding all four features. Overall speaking, the highest improvement over the baseline model 6.6% was achieved when adding the energy contour and speaker identity features. It seems that the speaker identity feature carries local spectro-temporal information of timbre such that it provides certain complementary information to the long-term energy contour feature. On the other hand, although the pitch contour feature provides a decent performance boost as shown in Table 4, it does not appear in the best combinations of features in Table 5. The observation is actually not surprising. Since both contours contain long-term information of timbre, the pitch contour feature probably does not provide complementary information to the energy contour feature. To sum up, these results indicate it is better to choose proper perceptual features for the proposed multi-objective SE model rather than adding as many features as possible.

Table 4: *The average improvement of all evaluation scores over the baseline model (BM).*

	Improvement (%)
BM + Speaker Identity	3.25
BM + Pitch	4.5
BM + Pitch Contour	5.75
BM + Energy Contour	6.5

Table 5: *Averaged evaluation scores of Noisy, Baseline model, and the best combinations of different numbers of features. The "Imp" is short for improvement.*

	PESQ	ESTOI	CSIG	COVL	Imp(%)
Noisy	1.4	0.48	1.48	1.25	–
Baseline model	1.72	0.47	2.17	1.58	–
+Energy Contour	1.89	0.5	2.25	1.67	6.5
+Energy Contour **+Speaker Identity**	1.88	0.5	2.27	1.67	6.6
+Energy Contour **+Speaker Identity** **+Pitch**	1.86	0.5	2.31	1.69	6.4
+All	1.83	0.49	2.28	1.68	5.25

We propose the multi-objective SE model by adding perceptual features as additional objectives. The features are preselected based on domain knowledge. For comparison, we also trained and tested a transformer based SE model using the same dataset. The PESQ, ESTOI, CSIG, and COVL scores of the transformer based SE model are 1.97, 0.52, 2.33, and 1.72, respectively. And its average performance improvement over the baseline SE model is 10.5%. Although its average performance gain 10.5% is higher than the average performance gain of the proposed model 6.6% (as shown in Table 5), its model size is much larger than the size of the proposed model (9.04 M vs 2.64 M in terms of parameter numbers). In other words, the transformer based SE model outperforms the proposed multi-objective SE model but at the price of with a much higher computational load.

5. Conclusion

We proposed a multi-objective SE model by adding perceptual characteristics based feature losses when updating the conventional SE model. Experimental results demonstrate that the proposed multi-objective model can preserve more perceptual characteristics when synthesizing speech and produce higher scores in all evaluation metrics than the baseline model. In addition, incorporating each of the features or combinations of these features could also improve performance of the baseline model. Therefore, we conclude that these perceptual characteristics based features are very helpful for SE. In future, we plan to evaluate the idea for time-domain SE by developing feature models in the time domain to combine with time-domain SE model using waveform loss. Also, we will investigate the benefit of using these perceptual characteristics based features in other speech-related tasks, such as speech separation.

6. References

[1] D. Wang, "Deep learning reinvents the hearing aid," *IEEE spectrum*, vol. 54, no. 3, pp. 32–37, 2017.

[2] Y.-H. Lai, F. Chen, S.-S. Wang, X. Lu, Y. Tsao, and C.-H. Lee, "A deep denoising autoencoder approach to improving the intelligibility of vocoded speech in cochlear implant simulation," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 7, pp. 1568–1578, 2016.

[3] M. Kolboek, Z.-H. Tan, and J. Jensen, "Speech enhancement using long short-term memory based recurrent neural networks for noise robust speaker verification," in *2016 IEEE spoken language technology workshop (SLT)*. IEEE, 2016, pp. 305–311.

[4] S. Shon, H. Tang, and J. Glass, "Voiceid loss: Speech enhancement for speaker verification," *arXiv preprint arXiv:1904.03601*, 2019.

[5] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "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, 2015, pp. 91–99.

[6] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745–777, 2014.

[7] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2007.

[8] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 27, no. 2, pp. 113–120, 1979.

[9] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[10] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE transactions on acoustics, speech, and signal processing*, vol. 33, no. 2, pp. 443–445, 1985.

[11] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder." in *Interspeech*, vol. 2013, 2013, pp. 436–440.

[12] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal processing letters*, vol. 21, no. 1, pp. 65–68, 2013.

[13] ——, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2014.

[14] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement," *arXiv preprint arXiv:2008.00264*, 2020.

[15] A. Li, W. Liu, X. Luo, C. Zheng, and X. Li, "Icassp 2021 deep noise suppression challenge: Decoupling magnitude and phase optimization with a two-stage deep network," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6628–6632.

[16] A. Pandey and D. Wang, "Tcnn: Temporal convolutional neural network for real-time speech enhancement in the time domain," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6875–6879.

[17] A. Defossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," *arXiv preprint arXiv:2006.12847*, 2020.

[18] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[19] K. Qian, Y. Zhang, S. Chang, M. Hasegawa-Johnson, and D. Cox, "Unsupervised speech decomposition via triple information bottleneck," in *International Conference on Machine Learning*. PMLR, 2020, pp. 7836–7846.

[20] B. Sisman and H. Li, "Wavelet analysis of speaker dependent and independent prosody for voice conversion." in *Interspeech*, 2018, pp. 52–56.

[21] C.-J. Peng, Y.-J. Chan, C. Yu, S.-S. Wang, Y. Tsao, and T.-S. Chi, "Attention-based multi-task learning for speech-enhancement and speaker-identification in multi-speaker dialogue scenario," in *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2021, pp. 1–5.

[22] F.-K. Chuang, S.-S. Wang, J.-w. Hung, Y. Tsao, and S.-H. Fang, "Speaker-aware deep denoising autoencoder with embedded speaker identity for speech enhancement." in *Interspeech*, 2019, pp. 3173–3177.

[23] Y.-J. Lu, C.-Y. Chang, Y. Tsao, and J.-w. Hung, "Speech enhancement guided by contextual articulatory information," *arXiv preprint arXiv:2011.07442*, 2020.

[24] T.-A. Hsieh, C. Yu, S.-W. Fu, X. Lu, and Y. Tsao, "Improving perceptual quality by phone-fortified perceptual loss for speech enhancement," *arXiv preprint arXiv:2010.15174*, 2020.

[25] A. Łańcucki, "Fastpitch: Parallel text-to-speech with pitch prediction," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6588–6592.

[26] M. Huang, "Development of taiwan mandarin hearing in noise test," *Department of speech language pathology and audiology, National Taipei University of Nursing and Health science*, 2005.

[27] A. S. Bregman, *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1994.

[28] B. C. Moore, *An introduction to the psychology of hearing*. Brill, 2012.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[30] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.

[31] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[32] 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, 2007.