TOWARDS AN ASR APPROACH USING ACOUSTIC AND LANGUAGE MODELS FOR SPEECH ENHANCEMENT

Khandokar Md. Nayem and Donald S. Williamson

Department of Computer Science, Indiana University, USA
knayem@iu.edu, *williads@indiana.edu*

ABSTRACT

Recent work has shown that deep-learning based speech enhancement performs best when a time-frequency mask is estimated. Unlike speech, these masks have a small range of values that better facilitate regression-based learning. The question remains whether neural-network based speech estimation should be treated as a regression problem. In this work, we propose to modify the speech estimation process, by treating speech enhancement as a classification problem in an ASR-style manner. More specifically, we propose a quantized speech prediction model that classifies speech spectra into a corresponding quantized class. We then train and apply a language-style model that learns the transition probabilities of the quantized classes to ensure more realistic speech spectra. We compare our approach against time-frequency masking approaches, and the results show that our quantized spectra approach leads to improvements.

Index Terms— speech enhancement, language model, speech quantization, deep learning

1. INTRODUCTION

Monaural speech enhancement is a challenging problem that aims to remove unwanted noise from a speech signal. The increasing usage of electronic devices, such as smart speakers, voice-controlled devices, and hearing aids increases the need for improved speech enhancement. Advancements in deep learning have led the field towards a solution, but, poor performance and unwanted distortions in noisy conditions require further improvements.

Speech enhancement is divided into two forms, either mask-based or signal-based approximation. A time-frequency (T-F) mask is estimated in mask-based approaches and it filters unwanted noise. Early mask-based approaches estimate the ideal binary mask (IBM) [1] and ideal ratio mask (IRM) [2]. More recent approaches estimate the phase-sensitive mask (PSM) [3], complex ratio mask (cIRM) [4] or parametric complex-valued T-F mask [5] to enhance magnitude and phase. On the other hand, signal approximation can be done in either the time [6, 7] or the T-F domains [8].

Deep clustering, which forms a binary mask for multi-speaker separation, has also been proposed [9]. This approach clusters each T-F unit into one of many clusters, which corresponds to a sound source. Hence, source classification is performed. Early approaches have also performed classification-based enhancement [10, 11]. In [12], separate discrete T-F masks for magnitude and phase responses are estimated using softmax activations, where recurrent networks are used to capture temporal correlations. The ideal quantized mask (IQM) has also recently been proposed [13]. It shows that quantization of the IRM, by coding each T-F IRM value into one of a number of quantized bins, is a reasonable representation of the IRM as assessed by human listeners. This study, however, did not evaluate estimated versions of the IQM. These studies essentially show that estimating quantized or class values is beneficial. However, although recurrent networks capture temporal correlations, these approaches do not ensure that the resulting speech spectra exhibit realistic spectral- and temporal-fine structure that occurs within real speech signals from human sources. Our recent work strives to enforce spectral-fine structure by incorporating an intra-spectral recurrence layer [14, 15], but they do not address temporal-fine structure and do not ensure that human-like spectra is generated.

Automatic speech recognition (ASR) helps ensure that realistic text transcriptions are generated by applying language models on top of DNN-based acoustic models [16]. Motivated by this, we propose a signal-approximation approach that uses a recurrent network to estimate quantized T-F spectra values, where we subsequently apply a spectral model to generate more realistic (human-like) spectra across time and frequency. In other words, our quantized spectra estimation is analogous to acoustic modeling, and our spectral model is akin to a language model. Here, quantization refers to treating T-F speech estimation as a classification problem, where each T-F spectral value is assigned to one of many quantized classes. We conduct a listening study to show that quantized speech is not acoustically different from clean speech, according to human listeners. We propose two quantized spectral models (QSM) that learn the transition probabilities between the quantized classes of speech across time and frequency. Unlike prior approaches [3, 17], which either treat ASR as a

back-end component or that uses features from ASR-models to improve ASR performance, our proposed approach uses ASR-equivalent acoustic and language models for speech enhancement. To the best of our knowledge, deep-learning based quantized speech approximation with a T-F level quantized spectral model has not been investigated for monaural speech enhancement.

2. PROPOSED APPROACH

Let's define clean speech as s_t and background noise as n_t at time t. The mixture of clean speech and noise is denoted as, $m_t = s_t + n_t$. Using the short-time Fourier transform (STFT), T-F domain signal $S_{t,k}$ is computed from s_t at time t and frequency k, where $S_{t,k} = |S_{t,k}|e^{i\theta^S_{t,k}}$. Enhancement of the noisy speech magnitude $|M_{t,k}|$ produces estimated clean magnitude $|\hat{S}_{t,k}|$. In this approach, we enhance the magnitude response, but use the noisy phase $\theta^M_{t,k}$ to reproduce an enhanced speech signal.

2.1. Speech Quantization

$|S_{t,k}| \in \{0, \Re^+\}$ is unbounded and continuous valued. Here, a scaling function $\mathcal{C}_{[0,r]}(\cdot)$ is used to constrain the values within the range $[0, r]$. We constrain the amplitude of a signal by setting r to 100. A quantization function $\mathcal{Q}_\chi(\cdot)$ converts the range constrained magnitude spectrogram into $\mathcal{D}$ number of bins which are χ steps apart. This produces quantized speech, i.e. $|S_{t,k}|^q = \mathcal{Q}_\chi\big(\mathcal{C}_{[0,r]}(|S_{t,k}|)\big)$. Now, a speech enhancement system can learn a function $F^q_\phi(\cdot)$ that maps between noisy speech $|M_{t,k}|$ and quantized clean speech $|S_{t,k}|^q$, and this becomes a $\mathcal{D}$ class classification problem, $|\hat{S}_{t,k}|^q = F^q_\phi\Big(|M_{t,k}|\Big)$.

Choosing the quantization step χ is a crucial part of speech quantization, as it has to be small enough so that human listeners do not notice the difference between quantized and original speech. For determining the best χ, we conduct a listening study whose details are in the subsection 3.1. An example of the original clean and quantized clean magnitude spectra are shown in Fig. 1, where $\chi = 2$ for display purposes.

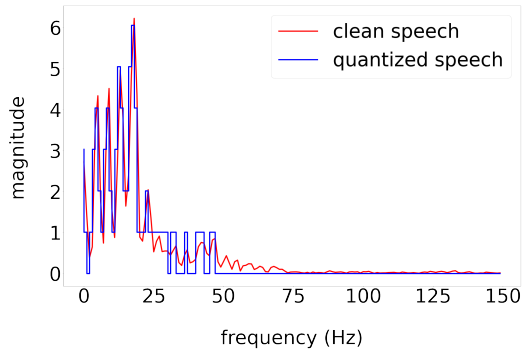

Fig. 1: Quantization of a clean magnitude spectrum.

2.2. Quantized Spectral Model

Traditionally, a language model (LM) is applied at the word or phoneme level, where the effectiveness of the LM depends on the text and its vocabulary. We propose an alternative view of a LM, where we consider each quantization level as a word. We consider bi-gram LM, which we refer to as the Quantized Spectral Model (QSM). Though we construct the QSM using quantized speech magnitudes from clean speech corpora, unlike traditional LMs, the QSM is less likely to suffer from the out of vocabulary problem when the model parameters, χ and r, are adequately defined.

We consider both mean and per-frequency-channel QSMs, computed along the time axis. For the mean QSM (mQSM), each entry refers to the transition probability between two time consecutive quantized levels, $mQSM = P(d_{t+1,:}|d_{t,:})$, where the probabilities are computed across all frequency channels. Similarly, the per-frequency QSM, fQSM, is defined as $fQSM_k = P(d_{t+1,k}|d_{t,k})$, where per-frequency transitions are stored. The probabilities are calculated by counting the level transitions, and then normalizing by the appropriate scalar. The mean QSM results in a single $\mathcal{D} \times \mathcal{D}$ transition probability matrix, whereas the per-frequency-channel QSM produces a $F \times \mathcal{D} \times \mathcal{D}$ probability matrix. F is the total number of frequency channels. To overcome the zero-probability problem in N-grams, we reevaluate the transition probabilities using Good-Turing smoothing [18].

2.3. Model architecture

We adopt a similar model structure as Chimera++ [19] for estimating the quantized speech value at each T-F point (Fig. 2). Multiple bi-direction LSTM (BLSTM) layers are applied to learn a T-F embedding for the inputted speech. In the output layer, we use a Y-shaped structure with two branches. The rightmost branch predicts the quantized class probability for the t-th time frame using a linear and softmax layer. This branch of the network has two losses, a cross-entropy loss to assess classification performance ($\mathcal{L}_{cls}$), and a regression loss, where the estimated expected quantized value is computed at each T-F bin and compared to the true quantized value, using the mean-square error. The regressed loss function is denoted as $\mathcal{L}_{reg}$. The $\mathcal{L}_{reg}$ is should help with distinguishing between between nearby classes. This is later shown in Table 1.

The left branch of our model performs deep clustering to separate speech from noise, and it serves as a regularizing term for this approach. The network computes $\mathcal{E}$ dimensional unit-length embedding vector $v_{t,k} \in \Re^{(1\times\mathcal{E})}$ corresponding to the t,k-th element in the input. Similarly, $y_{t,k} \in \Re^{(1\times\mathcal{G})}$ is a one-hot label vector indicating which source in a mixture dominates time-frequency bin (t,k). Since we have only speech and noise as sources, $\mathcal{G}$ is 2. Now, stacking these values, we form the embedding matrix $V \in \Re^{(TF\times\mathcal{E})}$, and the source label matrix $Y \in \Re^{(TF\times\mathcal{G})}$. The embedding V is

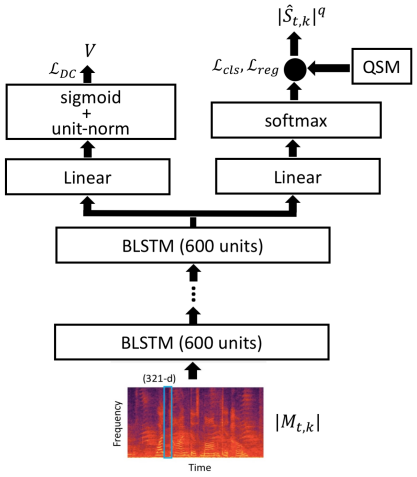

Fig. 2: Proposed network for speech enhancement.

learned by minimizing the following objective function:

$$\mathcal{L}_{DC} = \|VV^T - YY^T\|^2 \quad (1)$$

$$= \|V^TV\|^2 - 2\|V^TY\| + \|Y^TY\| \quad (2)$$

The overall loss function of our network with hyper-parameters λ_1 and λ_2 is defined as:

$$\mathcal{L} = (1-\lambda_1)\mathcal{L}_{DC} + \lambda_1\lambda_2\mathcal{L}_{cls} + \lambda_1(1-\lambda_2)\mathcal{L}_{reg} \quad (3)$$

This network predicts the quantization sequence conditioned on both the class probability and transition probability. QSM is trained separately and remains frozen when networks weights are updated during backpropagation. Then the enhanced speech sequence $|S_{1:T,:}|^q$ is of the optimal quantized class sequence which is calculated using:

$$|\hat{S}_{1:T,:}|^q = \underset{d_{1,:},\cdots,d_{T,:}}{\operatorname{argmax}} \prod_{i=1}^{T} P(M_{i,:}|d_{i,:})P(d_{i,:}|d_{i-1,:}) \quad (4)$$

Using a beam search algorithm, we can solve equation (4) and find the best quantized class d for each $|S_{t,k}|^q$.

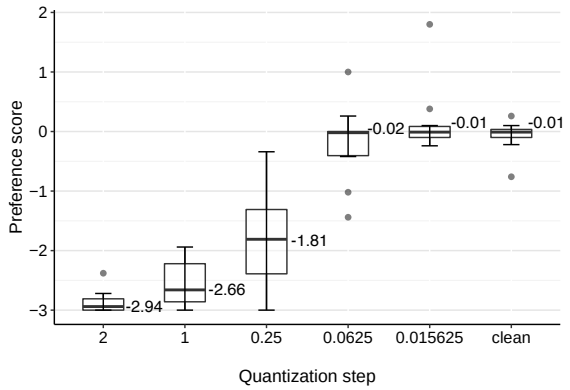

Fig. 3: Preference score for different quantization step from listeners study. Reference audio is clean speech.

3. EXPERIMENTS AND RESULTS

3.1. Listening study to determine quantization stepsize

We conduct an IRB-approved listening study using Amazon Mechanical Turk to determine the best quantization level as assessed by normal-hearing listeners. The sound quality is assessed at different quantization levels, where five values for χ, e.g. 2, 1, 0.25, 0.0625, 0.015625 are separately compared to the clean speech signal. Note that smaller χ results in more quantization levels ($\mathcal{D}$), which leads to better sound quality. These quantization levels result in quantized speech with equivalent signal to quantized-noise ratios (SQNR) of 14.21 dB, 17.78 dB, 26.5 dB, 36.25 dB, and 46.93 dB, respectively. The study is conducted as follows, the participant will listen to two audio signals, one is quantized and the other is clean audio. Then they provide a preference score using a Likert scale. The scale ranges from -3 to $+3$, where -3 refers to a strong preference towards the first signal, $+3$ refers to a strong preference towards the second signal, and 0 refers to no preference. Before providing a score, the participant can listen to the signals as many as times they like, where the scores are not limited to integer values. The two signals and corresponding quantization levels are randomly selected, and the participant listens to different audio clips in each question. The audio clips are chosen from the TIMIT corpus [20] spoken by both males and females in equal proportion.

The study session contains total 30 questions, which is preceded by a practice session of 7 questions. Ten participants (9 male, 1 female) who are native English speakers over the age of 18 participated, where a headset/headphone was required to be worn. On average, participants took 14 minutes to complete the study, they were given \$3 monetary incentive.

The results of the study are shown in Figure 3. For quantization levels 2, 1, and 0.25, negative scores indicate that these produce noticeably poorer sound quality. However, for $\chi = 0.0625$, the preference score is very close to 0, which means it is quite competitive with clean speech. Previous studies show that speech with SNRs$\geq$20 dB achieve sufficiently good perceptual quality [21] and intelligibility [22]. This is also the case in our study when $\chi = 0.0625$ (e.g.$\approx$36.25 dB SQNR). Therefore, we choose $\chi = 0.0625$ for quantization, which results in 1600 quantization classes.

3.2. Experimental setup and results

We train our model on the IEEE and TIMIT speech corpora. The IEEE corpus consists of 720 utterances from a single male speaker, and the TIMIT corpus has 6300 utterances from multiple male and female speakers. Our proposed QSM is trained on the clean speech of both these datasets.

Three non-overlapping sets of 50, 11, and 18.3 hrs are developed for the training, cross-validation, and testing sets, respectively. The training and validation data is generated at -3, 0, and 3 dB signal-to-noise ratios (SNRs) using four noise

Table 1: Average scores for each approach. Best results are shown in **bold**.

	IEEE corpus			TIMIT corpus		
	PESQ	SI-SDR	ESTOI	PESQ	SI-SDR	ESTOI
Mixture	1.86	1.8	0.53	1.81	-2.57	0.5
Chi++$_{\text{IQM2}}$	2.18	0.34	0.64	2.06	0.4	0.6
Chi++$_{\text{QM3}}$	2.25	0.41	0.68	2.08	0.43	0.64
Chi++$_{\text{IQM4}}$	2.32	0.63	0.71	2.14	0.52	0.68
Chi++$_{\text{IQM8}}$	2.37	0.72	0.73	2.1	0.53	0.69
Chimera [9]	2.4	0.81	0.75	2.16	0.49	0.69
Chi++$_{\text{tPSA}}$ [19]	2.46	0.84	0.76	2.25	0.74	0.72
Chi++$_{\text{quant}}$	2.44	0.82	0.75	2.2	0.63	0.67
Chi++$_{\text{mQSM,greedy}}$	2.45	0.88	0.8	2.26	0.81	0.74
Chi++$_{\text{fQSM,greedy}}$	2.46	0.93	0.82	2.27	0.84	0.74
Chi++$_{\text{mQSM,bS}}$	**2.48**	0.97	**0.83**	2.3	0.89	0.75
Chi++$_{\text{fQSM,bS}}$	**2.48**	**1.04**	**0.83**	**2.34**	**0.95**	**0.78**

types (speech-shaped noise, cafeteria, factory, and babble). We test with two additional SNRs (-6 and 6 dB), which are unseen by the recurrent model. All the signals are sampled at 16 kHz. The spectrogram is generated using a 640-point DFT with a Hann window of 40ms and a 20ms frameshift.

Our baseline network has four BLSTM layers of 600 cells with dropout layers between each of the BLSTM layers with dropout rate of 0.3. For the embedding vector, we use $\mathcal{E} = 20$. QSM is used in batch-wise on the model output to train the network with the loss function $\mathcal{L}$. The softmax is used as the activation function for the output layers that predict the quantized class. A sigmoid is used for the embedding approximation and gate activation function, while $\tanh$ functions are used for the cell and hidden states. Batch normalization is performed between each layer. Adam optimization is used with learning rate of 0.001. In the loss function, λ_1 and λ_2 are set to 0.5 and 0.975, respectively.

In our proposed approach, we use the same baseline network and investigate with different QSM techniques. We incorporate mQSM and fQSM which are denoted as Chi++$_{\text{mQSM}}$ and Chi++$_{\text{fQSM}}$ respectively. Also, we experiment with optimal quantization class sequence decoding algorithms. We try a greedy approach which assumes that quantized classes are pair-wise independent. The greedy approach is faster in decoding the optimal sequence. Additionally, we use an N-beam search approach, which does a beam search with the N best candidates.

We compare our method against the state-of-art models chimera [9] and chimera++ [19] which are mask-based approaches. We refer to these approaches as Chimera and Chi++$_{\text{tPSA}}$. Chimera predicts a magnitude mask and enhances only the magnitude of the mixture, however, Chi++$_{\text{tPSA}}$ estimates a phase-sensitive mask. Previous studies [23, 24] compare their models to Chimera models that enhance speech in non-speech noisy conditions with multi-talker speech. We compare against the Chimera networks that are trained for the speech enhancement task. To further investigate the effectiveness of quantized masking approaches, we use the Chimera++ model to estimate the IQM [13], where the models are Chi++$_{\text{IQM2}}$, Chi++$_{\text{IQM3}}$, Chi++$_{\text{IQM4}}$, and Chi++$_{\text{IQM8}}$ which predicts enhanced speech using IQM2, IQM3, IQM4, and IQM8 respectively. Here X in IQMX refers to the number of attenuation levels in the mask (see [13]). Additionally, we compare with a Chimera++ network that predicts quantized speech Chi++$_{\text{quant}}$ (proposed without the QSM). All the approaches are evaluated with three commonly-used objective metrics, namely, the PESQ, the scale-invariant SDR (SI-SDR) [25], and the extended STOI (ESTOI) [26].

We compare the performance scores of all the models in both seen (-3, 0, 3 dB) and unseen (-6, 6 dB) SNR conditions for IEEE and TIMIT data corpus. Table 1 shows the average scores of all the models. Our proposed Chi++$_{\text{fQSM}}$ outperforms all the other models with per-frequency QSM information and Chi++$_{\text{mQSM}}$ closely follows. Comparing with the best performing mask-based approach Chi++$_{\text{tPSA}}$, both Chi++$_{\text{fQSM,bs}}$ and Chi++$_{\text{mQSM,bS}}$ gives 0.2 PESQ gain and 0.07 ESTOI gain using the IEEE corpus. For TIMIT, Chi++$_{\text{fQSM,bs}}$ outperforms Chi++$_{\text{tPSA}}$ by 0.21 according to SI-SDR. Additionally, Chi++$_{\text{quant}}$ exceeds the Chimera model in all performance metrics; but still has lower scores than QSM infused models which indicates the advantages of a spectral model. It is worth noting that the proposed models outperform Chi++$_{\text{tPSA}}$, even though this approach enhances the magnitude and phase responses of a noisy speech signal, whereas our approaches only enhance the magnitude responses. It is likely that further gains will occur if our approach additionally enhances phase. Additionally, performance when estimating a quantized mask (e.g. IQM) tends to be lower than the phase-based masking approaches and proposed approaches. This occurs when the spectral model is and is not used. This indicates that quantizing the spectra leads to better performance than quantizing a mask. This likely occurs because the mask values are already in a small and constrained range.

For optimal quantized class decoding, we use both greedy and N-beam search (bS) approaches. Beam searching shows superiority likely because it covers a bigger search space of $(N \times T)$ class candidates ($N = 100$). With the beam search algorithm, Chi++$_{\text{fQSM,bS}}$ gains 0.9 more SI-SDR points for IEEE than its greedy version. Also with TIMIT, Chi++$_{\text{fQSM,bS}}$ receives a 0.7 PESQ gain compared to Chi++$_{\text{fQSM,greedy}}$.

4. CONCLUSION

Our proposed quantized speech classification approach with an ASR-style language model successfully enhances the speech mixture and outperforms T-F masking-based approaches. It shows that signal-approximation can be done successfully if the appropriate training target is considered. This approach, however, considers only bi-gram spectral models which are generated by considering only along-time transitions. In the future, we will explore higher-order N-gram models that consider both temporal and spectral transitions to enhance both magnitude and phase responses.

5. REFERENCES

[1] N. Li and P. C. Loizou, "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," *The JASA*, vol. 123, pp. 1673–1682, 2008.

[2] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proc. ICASSP*, pp. 7092–7096, 2013.

[3] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. ICASSP*, pp. 708–712, 2015.

[4] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM TASLP*, vol. 24, pp. 483–492, 2015.

[5] J. Lee and H.-G. Kang, "A joint learning algorithm for complex-valued tf masks in deep learning-based single-channel speech enhancement systems," *IEEE/ACM TASLP*, vol. 27, pp. 1098–1108, 2019.

[6] Y. Luo and N. Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," in *Proc. ICASSP*, pp. 696–700, 2018.

[7] A. Pandey and D. Wang, "A new framework for cnn-based speech enhancement in the time domain," *IEEE/ACM TASLP*, vol. 27, pp. 1179–1188, 2019.

[8] B. O. Odelowo and D. V. Anderson, "A study of training targets for deep neural network-based speech enhancement using noise prediction," in *Proc. ICASSP*, pp. 5409–5413, 2018.

[9] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. ICASSP*, pp. 31–35, 2016.

[10] Y. Wang, K. Han, and D. Wang, "Exploring monaural features for classification-based speech segregation," *IEEE TASLP*, vol. 21, pp. 270–279, 2012.

[11] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation," *IEEE TASLP*, vol. 21, pp. 1381–1390, 2013.

[12] J. Le Roux, G. Wichern, S. Watanabe, A. Sarroff, and J. R. Hershey, "Phasebook and friends: Leveraging discrete representations for source separation," *IEEE JSTSP*, vol. 13, pp. 370–382, 2019.

[13] E. W. Healy and J. L. Vasko, "An ideal quantized mask to increase intelligibility and quality of speech in noise," *The JASA*, vol. 144, pp. 1392–1405, 2018.

[14] K. M. Nayem and D. S. Williamson, "Incorporating intra-spectral dependencies with a recurrent output layer for improved speech enhancement," in *IEEE MLSP*, pp. 1–6, 2019.

[15] K. M. Nayem and D. S. Williamson, "Monaural speech enhancement using intra-spectral recurrent layers in the magnitude and phase responses," in *Proc. ICASSP*, pp. 6224–6228, 2020.

[16] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *ICML*, pp. 1764–1772, 2014.

[17] 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 *LVA/ICA*, pp. 91–99, Springer, 2015.

[18] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, vol. 2. 2008.

[19] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Alternative objective functions for deep clustering," in *Proc. ICASSP*, pp. 686–690, 2018.

[20] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon technical report*, vol. 93, 1993.

[21] P. C. Wong, A. K. Uppunda, T. B. Parrish, and S. Dhar, "Cortical mechanisms of speech perception in noise," 2008.

[22] D. Wang and G. J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications*. Wiley-IEEE press, 2006.

[23] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. L. Roux, "Wham!: Extending speech separation to noisy environments," *arXiv preprint arXiv:1907.01160*, 2019.

[24] G.-P. Yang, C.-I. Tuan, H.-Y. Lee, and L.-s. Lee, "Improved speech separation with time-and-frequency cross-domain joint embedding and clustering," *arXiv preprint arXiv:1904.07845*, 2019.

[25] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?," in *Proc. ICASSP*, pp. 626–630, 2019.

[26] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM TASLP*, vol. 24, pp. 2009–2022, 2016.