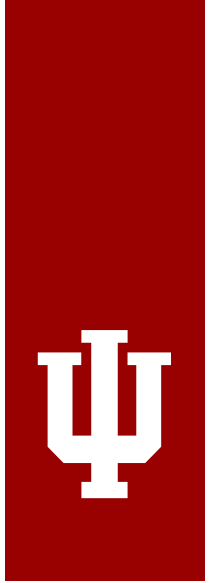


Towards an ASR Approach Using Acoustic and Language Models for Speech Enhancement

Khandokar Md. Nayem and Donald S. Williamson

Department of Computer Science, Indiana University, IN, USA

LUDDY
SCHOOL OF INFORMATICS,
COMPUTING, AND ENGINEERING

46th International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2021)

Introduction

Monaural speech enhancement is a challenging problem that aims to remove unwanted noise from a target speech signal.

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.

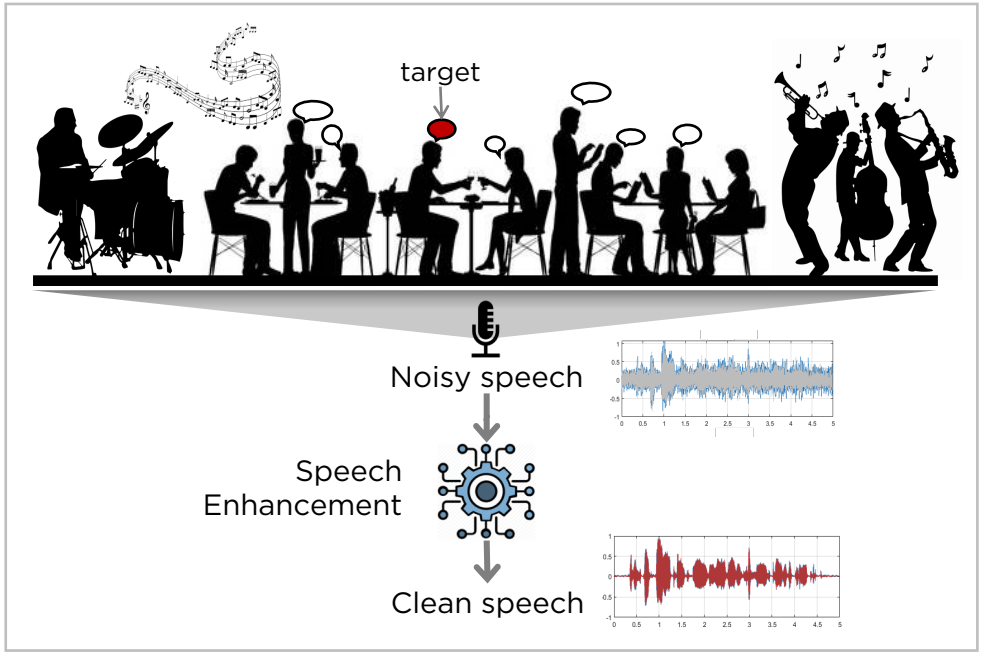


Image source, https://clipground.com/

Speech Enhancement Approaches

Mask-based speech enhancement

Li et al. 2008 ideal binary mask (IBM), Narayanan et al. 2013 ideal ratio mask (IRM), Erdogan et al. 2015 phase-sensitive mask (PSM), Williamson et al. 2015 complex ideal ratio mask (cIRM), Lee et al. 2019 parametric complex-valued time-frequency (TF) mask.

Signal approximation-based speech enhancement

Luo et al. 2018 and Wang et al. 2019 propose time domain end-to-end speech waveform enhancement, Odelowo et al. 2018 presents enhancement in TF domain.

Deep clustering

Hershey et al. 2016 suggest a binary mask for source separation task

Classification-based enhancement

Wang et al. 2012 and 2013 perform IBM estimation, Roux et al. 2019 propose discrete Codebook representation to pose the problem as a classification one.

ASR features in speech enhancement

Erdogan et al. 2015 and Weninger et al. 2015 utilize ASR feature as only input feature.

Motivation

Ideal Quantized Mask (IQM)

Healy et al. 2018 proposes four different quantized levels ideal masks.

Quantized ideal mask, however, does not consider spectral correlations along the frequency axis.

Language model in ASR

Graves et al. 2014 show the advantages of incorporating language model in end-to-end automatic speech recognition (ASR) task.

Language model (LM) can be a better way to incorporate linguistic property of speech in end-to-end speech approximation system.

Phonetic or textual LM heavily depend on rich and robust datasets, which is now commercially in practice, whereas spectral LM focuses on spectral property of human speech can be an alternate view.

Proposed approach

We propose a signal-approximation approach to estimate quantized T-F spectral values, where we subsequently apply a spectral model to generate more realistic speech.

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 .

Spectral language-style model that learns the transition probabilities of the quantized classes across time and frequency to ensure more realistic speech spectra.

Our quantized spectra estimation is analogous to acoustic modeling, and our spectral model is akin to a language model.

Problem Formulation

In the time domain, $m_t = s_t + n_t$

In the time-frequency (T-F) domain, $M_{t,k} = |M_{t,k}|e^{i\theta^M_{t,k}}$

$S_{t,k} = |S_{t,k}|e^{i\theta^S_{t,k}}$

$|S_{t,k}|$ → clean speech magnitude

$\theta^S_{t,k}$ → clean speech phase

$N_{t,k} = |N_{t,k}|e^{i\theta^N_{t,k}}$

$|N_{t,k}|$ → noise magnitude

$\theta^N_{t,k}$ → noise phase

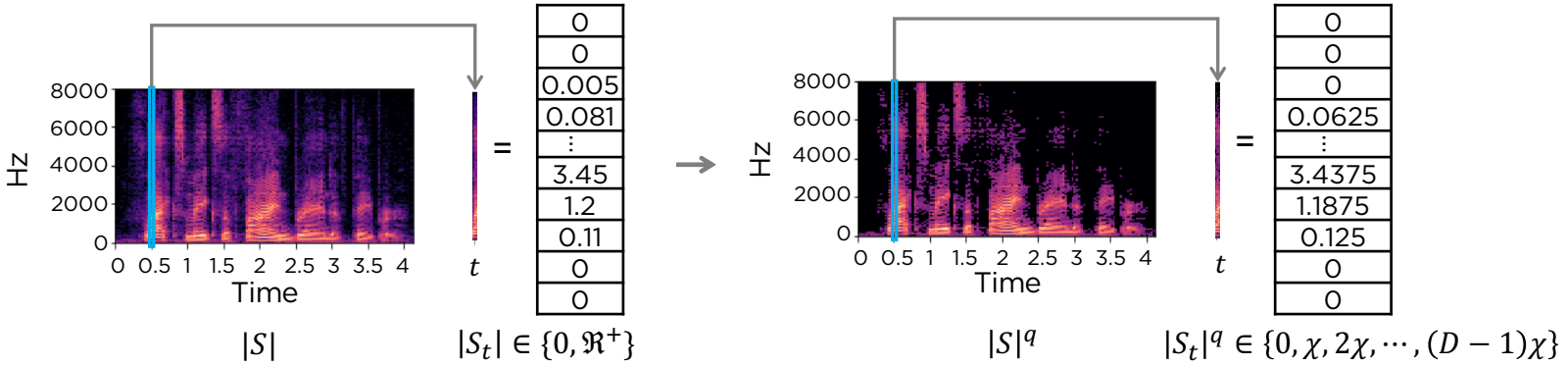
Estimate clean speech, $|\hat{S}_{t,k}| = F_\phi(M_{t,k}) = F_\phi(|M_{t,k}|, \theta^M_{t,k})$

m_t → noisy speech
s_t → clean speech
n_t → noise
t → time index

$M_{t,k}$ → T-F noisy speech
$|M_{t,k}|$ → magnitude response
$\theta^M_{t,k}$ → phase response
k → frequency index

ϕ → model parameter

Speech Quantization



Speech $|S_{t,k}| \in \{0, \Re^+\}$ is unbounded and continuous valued.

We constrain and quantize the speech applying a scaling function $\mathcal{C}_{[0,r]}(\cdot)$ and a quantization function $\mathcal{Q}_\chi(\cdot)$,

$$|S_{t,k}|^q = \mathcal{Q}_\chi\left(\mathcal{C}_{[0,r]}(|S_{t,k}|)\right), \text{ where } \mathcal{Q}_\chi(|S_{t,k}|) = \chi \cdot \underset{i}{\operatorname{argmin}}\left(\{0, \chi, 2\chi, \cdots, (D-1)\chi\} - |S_{t,k}|\right)$$

Here r is the max spectrum value and χ is the quantization step size which yields total $\mathcal{D}$ number of bins.

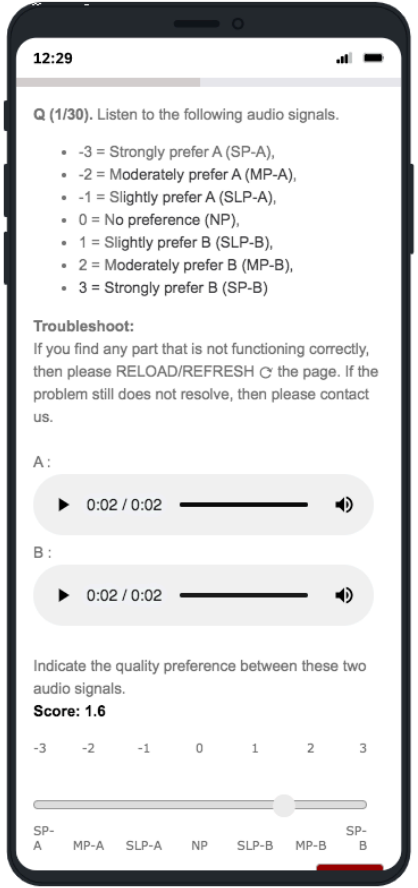
Listening Study

We conduct an IRB-approved listening study using Amazon Mechanical Turk to determine the best quantization level as assessed by normal-hearing listeners.

Five different quantization levels using step sizes of χ = 2, 1, 0.25, 0.0625, 0.015625 are separately compared to clean reference speech.

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 session contains a 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.



Listening Study

For quantization levels 2, 1, and 0.25, negative scores indicate that these produce noticeably poorer sound quality.

For χ = 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 ≥ 20 dB achieve sufficiently good perceptual quality [1] and intelligibility [2]. This is also the case in our study when χ = 0.0625 (e.g.,≈36.25 dB SQNR).

So, we choose $\chi = 0.0625$ for quantization, which results in 1600 quantization classes.

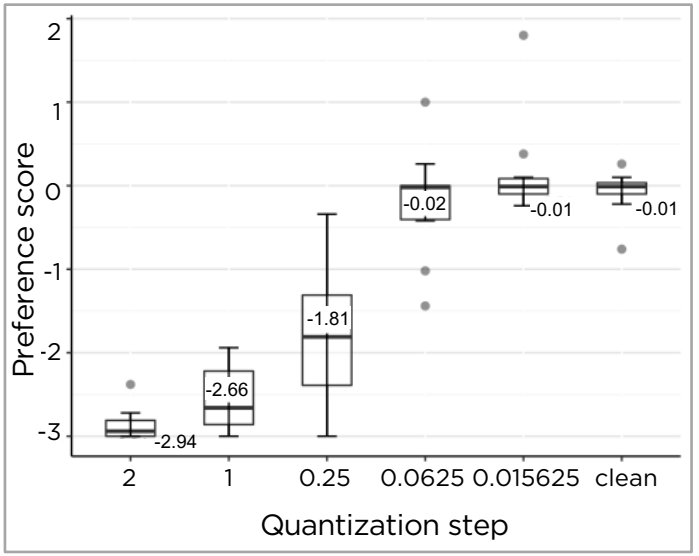


Fig. Preference score for different quantization step from listener study. Reference audio is clean speech.

[1] P. C. Wong, et al., "Cortical mechanisms of speech perception in noise," 2008.
[2] D. Wang, et al., Computational auditory scene analysis: Principles, algorithms, and applications. Wiley-IEEE press, 2006.

Speech Enhancement Model

We adopt a Chimera++ type model structure for estimating the quantized speech value at each T-F point.

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 ($\mathcal{L}_{reg}$), 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 (MSE).

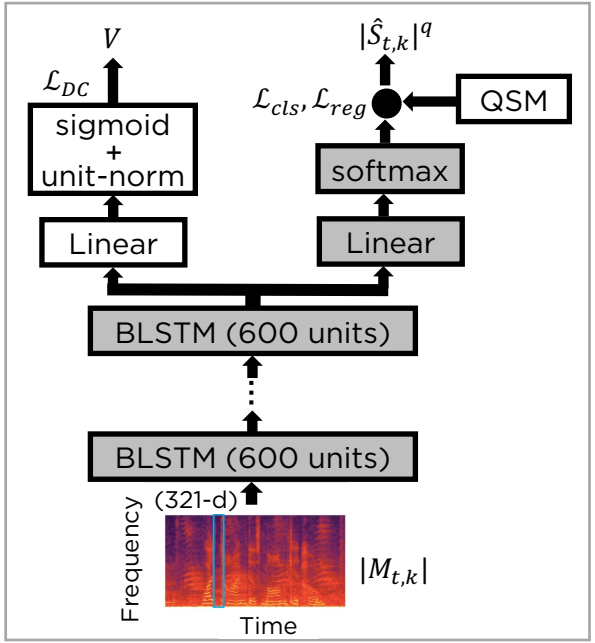


Fig. Proposed network for speech enhancement.

Speech Enhancement Model

The left branch of our model performs deep clustering to separate speech from noise, and it serves as a regularizing term for this approach.

We form the embedding matrix $\mathrm{V} \in \Re^{(TF \times \varepsilon)}$ and source label vector indicator matrix $\mathrm{Y} \in \Re^{(TF \times \mathcal{G})}$.

The source label matrix V is learned by minimizing the following objective function:

$$\mathcal{L}_{DC} = \|VV^T - YY^T\|^2 = \|V^TV\|^2 - 2\|V^TY\|^2 + \|Y^TY\|^2$$

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

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

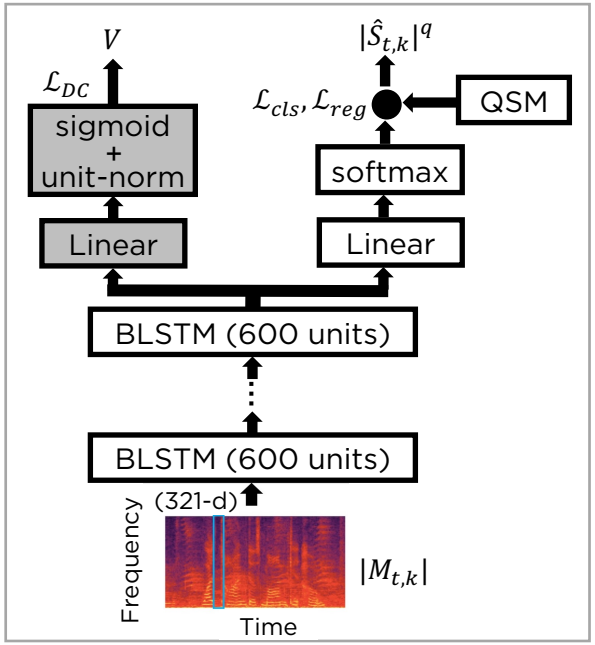
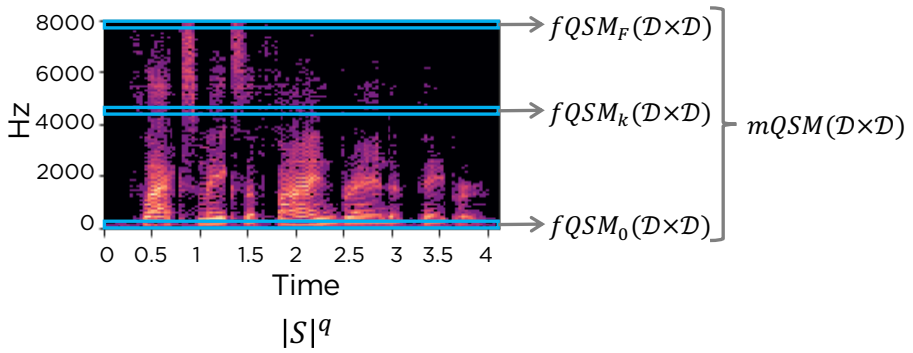


Fig. Proposed network for speech enhancement.

Quantized Spectral Model

Traditional 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).



We consider per-frequency QSM (fQSM) where the probabilities are computed per-frequency channel and each entry (d) refers to the transition probability between two-time consecutive quantized levels.

$$fQSM_k = P(d_{t+1,k}|d_{t,k})$$

Mean QSM (mQSM) is defined across all frequency channels.

$$mQSM = P(d_{t+1,:}|d_{t,:})$$

SE+QSM Model

This network predicts the quantization sequence conditioned on both the class probability and transition probability.

QSM is trained separately and remains frozen when network weights are updated during backpropagation.

The enhanced speech sequence $|S_{1:T,:}|^q$ is of the optimal quantized class sequence which is calculated using:

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

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

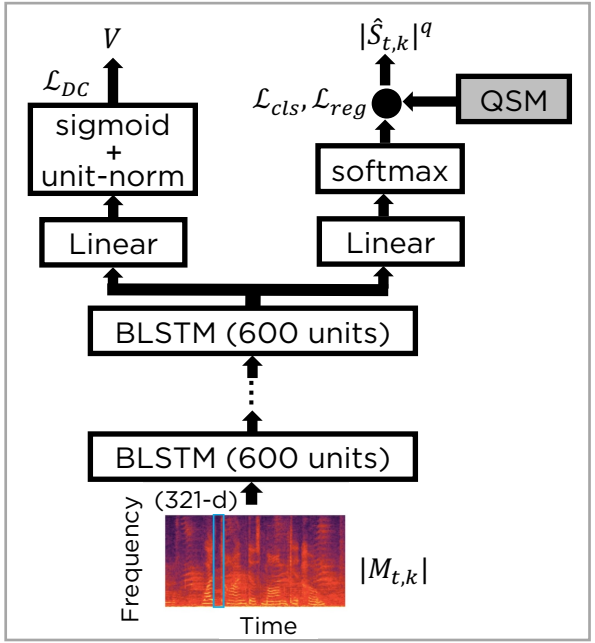


Fig. Proposed network for speech enhancement.

Dataset

Train and evaluate using IEEE male (single speaker, 720 utterances) and TIMIT (multiple speaker, 6300 utterances) speech corpora.

Our proposed QSM is trained on the clean speech of both these datasets.

4 Noise types- speech-shaped noise (SSN), cafeteria, factory, and babble.

Trained and validated in 3 SNR levels (-3, 0, 3 dB) and tested in additional 2 SNR levels (-6 and 6 dB).

Comparison approaches

Chimera model → Chimera [1]

Chimera++ model → Chi++$_{\text{tPSA}}$ [2]

Different IQMs [3] estimated Chi++ models → Chi++$_{\text{IQMX}}$

Quantized speech estimated Chi++ model → Chi++$_{\text{quant}}$

Proposed Chi++ models with QSM → Chi++$_{\text{mQSM/fQSM,greedy/bS}}$

[1] J. R. Hershey, et al., "Deep clustering: Discriminative embeddings for segmentation and separation," in Proc. ICASSP, pp. 31–35, 2016.
[2] Z.-Q. Wang, et al., "Alternative objective functions for deep clustering," in Proc. ICASSP, pp. 686–690, 2018.
[3] E. W. Healy, et al., "An ideal quantized mask to increase intelligibility and quality of speech in noise," *JASA*, pp. 1392–1405, 2018.

Experimental Setup

Baseline network has four LSTM layers with 600 units.

Dropout layers between each BLSTM layers with rate 0.3.

Softmax and sigmoid activation functions are used for the output layers that predict quantized class and embedding approximation, respectively.

For the embedding vector, we use $\varepsilon = 20$ and $\mathcal{G} = 2$.

Adam optimizer with learning rate 0.001, early stopping by validation set.

Loss function parameters, $\lambda_1 = 0.5$ and $\lambda_1 = 0.975$

Results

Table: 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++$_{\text{IQM3}}$	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 [1]	2.4	0.81	0.75	2.16	0.49	0.69
Chi++$_{\text{tPSA}}$ [2]	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**

[1] J. R. Hershey, et al., "Deep clustering: Discriminative embeddings for segmentation and separation," in Proc. ICASSP, pp. 31–35, 2016.
[2] Z.-Q. Wang, et al., "Alternative objective functions for deep clustering," in Proc. ICASSP, pp. 686–690, 2018.

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.

Thank You

Khandokar Md. Nayem
knayem@iu.edu

Donald S. Williamson
williads@indiana.edu

Department of Computer Science
Audio, Speech and Information Retrieval (ASPIRE) lab
https://aspire.sice.indiana.edu/
Indiana University, IN, USA