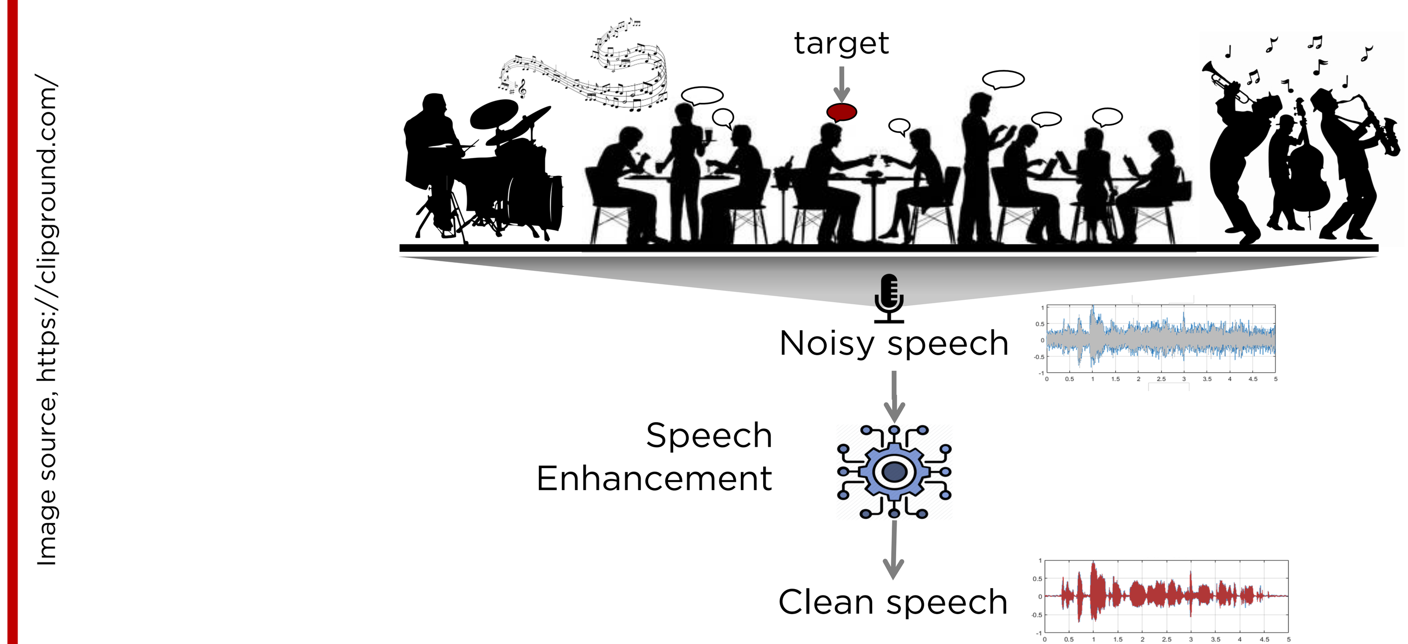
Towards an ASR Approach Using Acoustic and Language Models for Speech Enhancement

Khandokar Md. Nayem and Donald S. Williamson
Department of Computer Science, Indiana University, IN, USA
knayem@iu.edu, williads@Indiana.edu

Speech Enhancement

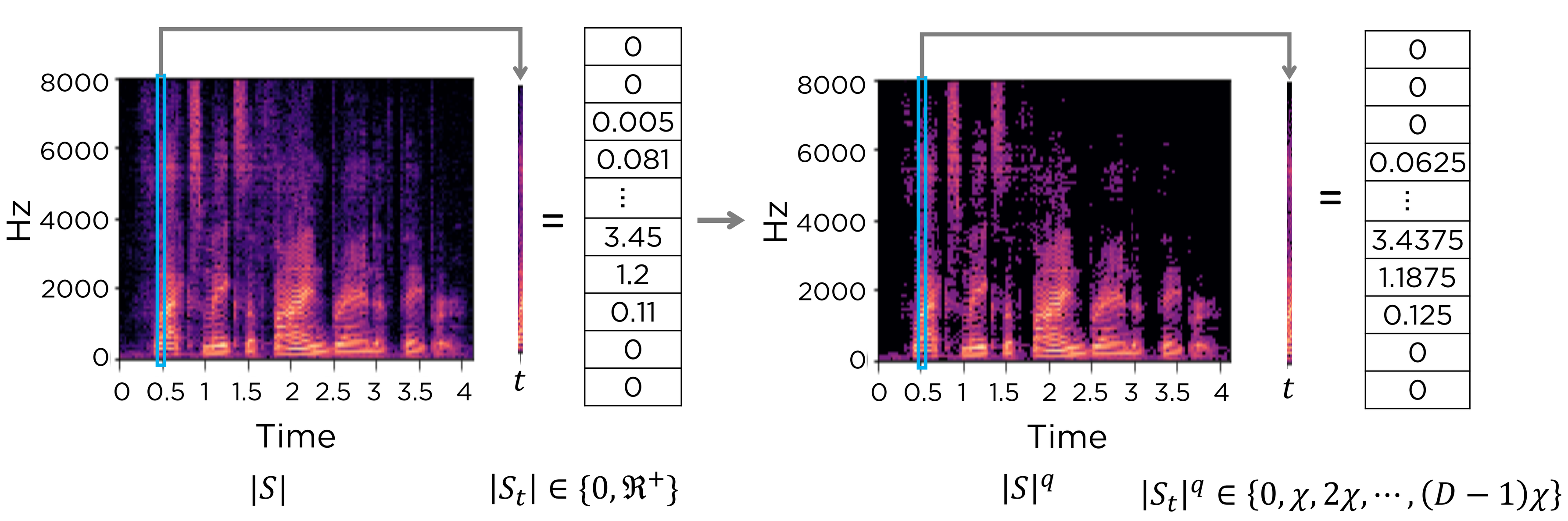


- Monaural speech enhancement (SE) is a challenging problem that aims to remove unwanted noise from a target speech signal.
- Increasing usage of electronic devices increases the need for improved speech enhancement.
- Poor performance and unwanted distortions in noisy conditions require further improvements.

Motivation

- Healy et al. 2018 propose the ideal quantized mask (IQM), which converts SE from a regression problem to a classification problem.
- The quantized mask, however, does not consider spectral correlations along the frequency axis.
- A language model (LM) can be a better way to incorporate linguistic property of speech in an end-to-end speech approximation system.
- Hence, a spectral LM which focuses on spectral properties of human speech can be an alternate approach.

Quantized Spectral Model



- We constrain and quantize the unbounded continuous valued speech $|S_{t,k}| \in \{0, \Re^+\}$ by,

$$|S_{t,k}|^q = \mathcal{Q}_\chi(\mathcal{C}_{[0,r]}(|S_{t,k}|))$$

$$\mathcal{Q}_\chi(|S_{t,k}|) = \chi \cdot \underset{i}{\operatorname{argmin}}(\{0, \chi, 2\chi, \cdots, (\mathcal{D}-1)\chi\} - |S_{t,k}|)$$

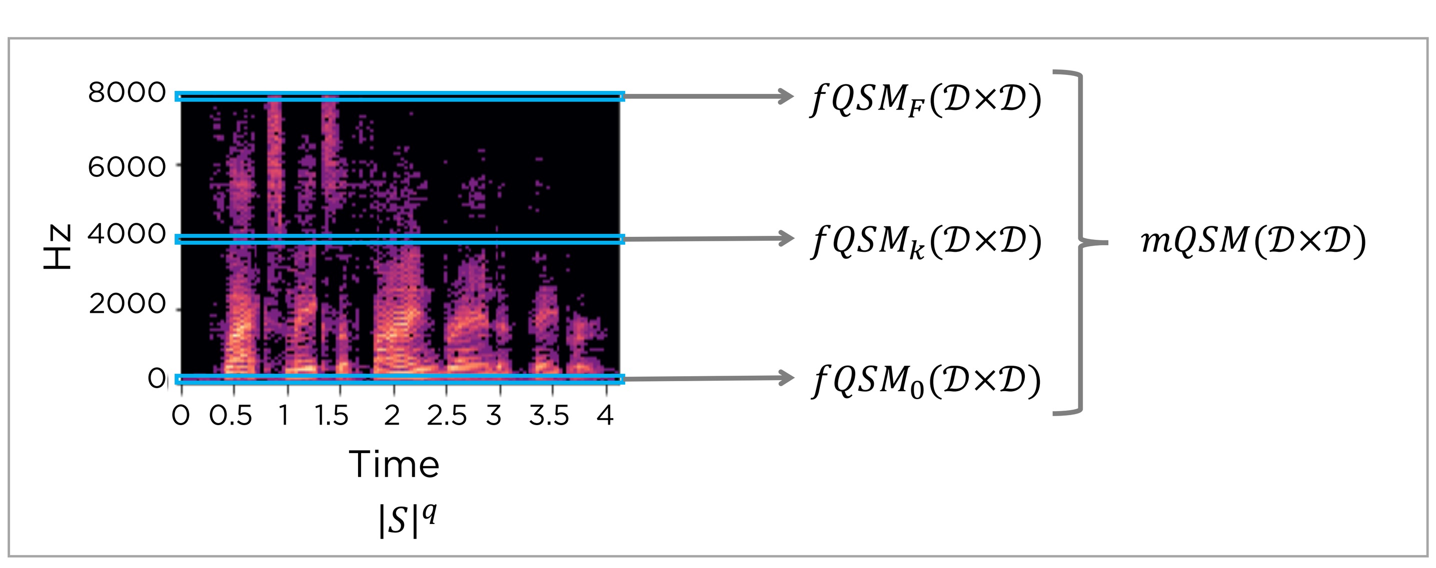
where r is the max spectrum value, χ is the quantization step size, $\mathcal{C}_{[0,r]}(\cdot)$ is a scaling function, $\mathcal{Q}_\chi(\cdot)$ is a quantization function which converts the range constrained magnitude spectrogram into total $\mathcal{D}$ $\left(=\frac{r}{\chi}\right)$ number of bins.

- Unlike traditional word or phoneme level-based language model (LM), 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 a mean QSM (mQSM) where the probabilities are computed across all frequency channels and each entry (d) refers to the transition probability between two-time consecutive quantized levels.

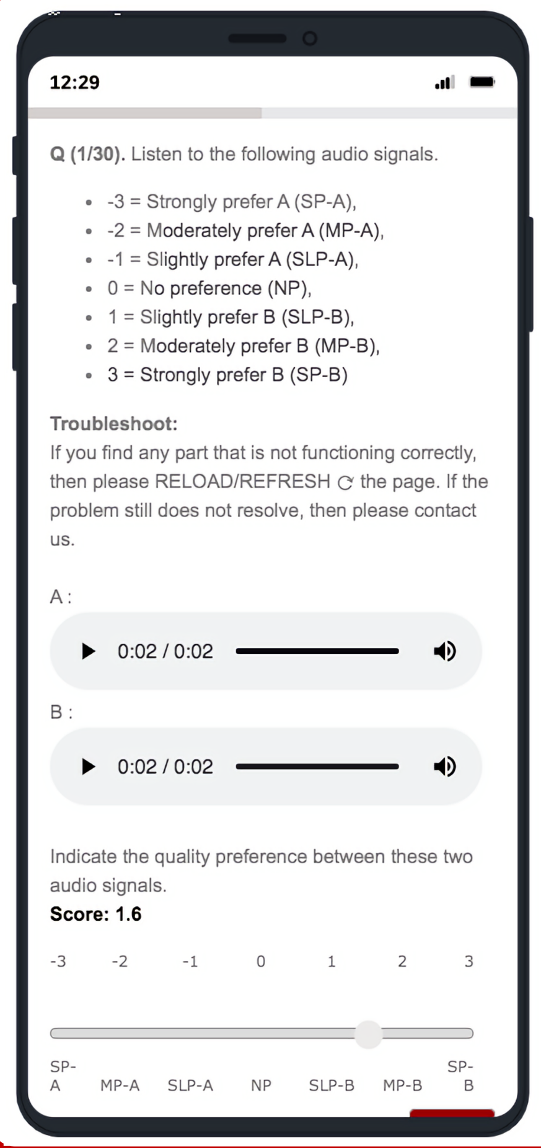
$$mQSM = P(d_{t+1,:}|d_{t,:})$$

- Per-frequency QSM (fQSM) is defined per-frequency transitions are stored.

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



Listening Study



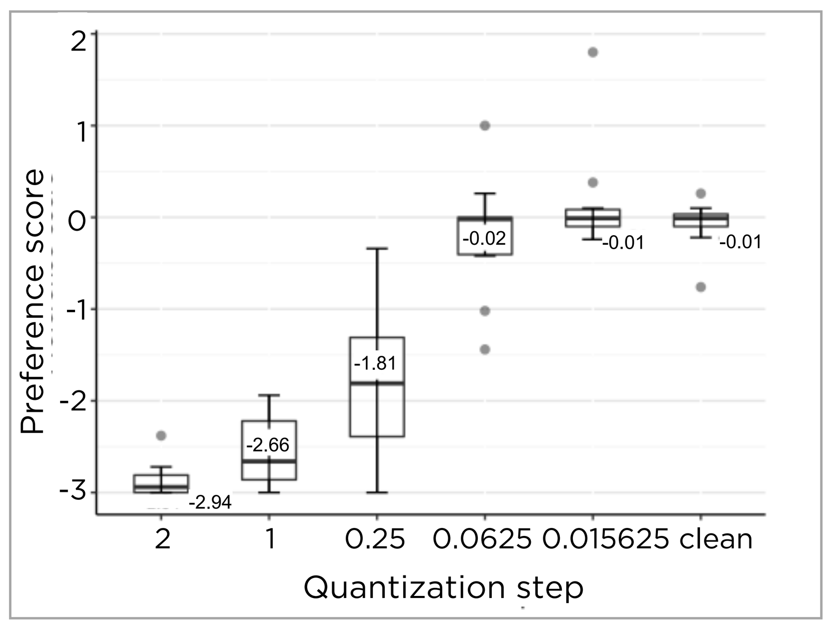


Fig. Preference score for different quantization step from listener study.

- We conduct an IRB-approved listening study using Amazon Mechanical Turk, where 5 different quantization levels are assessed.
- The study session consists of 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.

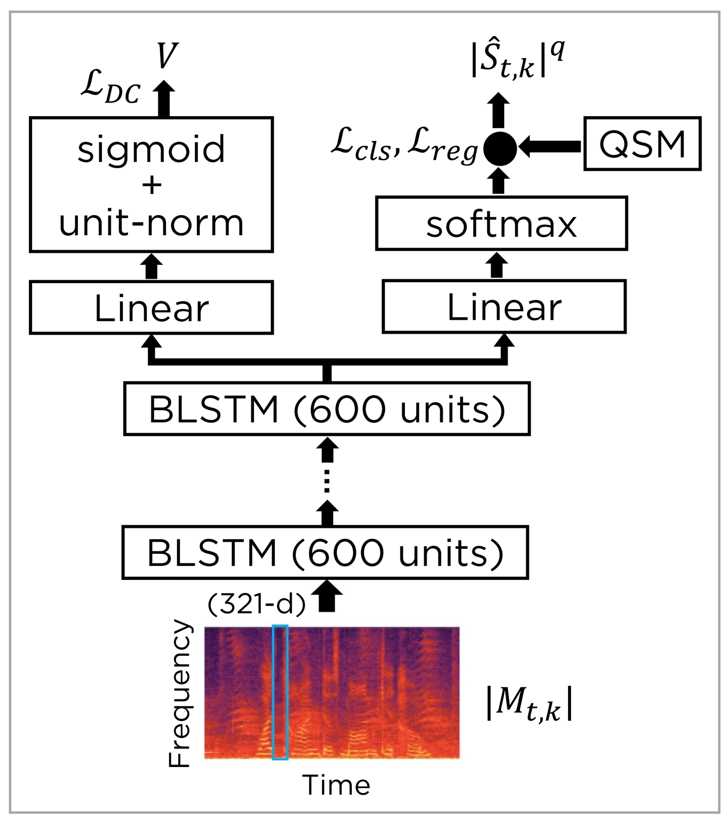
Proposed Speech Enhancement Model

- Rightmost branch predicts the quantized class probability for the t-th time frame with two losses, a cross-entropy loss ($\mathcal{L}_{cls}$), and a regression loss ($\mathcal{L}_{reg}$).
- Left branch performs deep clustering with ($\mathcal{L}_{DC}$) to separate speech from noise.
- Loss function is defined as,

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

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

where V is embedding matrix, Y is source hot-vector, and λ_1 and λ_2 are hyper-parameters.



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

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

Experiments and Results

- Train and evaluate using IEEE male (single speaker, 720 utterances) and TIMIT (multiple speakers, 6300 utterances) speech corpora.
- Noise types: speech-shaped noise (SSN), cafeteria, factory, and babble.
- Trained in 3 SNR levels (-3, 0, 3 dB), tested in additional 2 SNR levels (-6, 6 dB).

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
$Chi{+}{+}_{IQM2}$	2.18	0.34	0.64	2.06	0.4	0.6
$Chi{+}{+}_{IQM3}$	2.25	0.41	0.68	2.08	0.43	0.64
$Chi{+}{+}_{IQM4}$	2.32	0.63	0.71	2.14	0.52	0.68
$Chi{+}{+}_{IQM8}$	2.37	0.72	0.73	2.1	0.53	0.69
Chimera	2.4	0.81	0.75	2.16	0.49	0.69
$Chi{+}{+}_{tPSA}$	2.46	0.84	0.76	2.25	0.74	0.72
$Chi{+}{+}_{quant}$	2.44	0.82	0.75	2.2	0.63	0.67
$Chi{+}{+}_{mQSM,greedy}$	2.45	0.88	0.8	2.26	0.81	0.74
$Chi{+}{+}_{fQSM,greedy}$	2.46	0.93	0.82	2.27	0.84	0.74
$Chi{+}{+}_{mQSM,bS}$	**2.48**	0.97	**0.83**	2.3	0.89	0.75
$Chi{+}{+}_{mQSM,bS}$	**2.48**	**1.04**	**0.83**	**2.34**	**0.95**	**0.78**

Conclusion and Future Work

- Improvements in a variety of noises and SNR values prove that 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 quantized 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.