SPECTRAL FEATURE MAPPING WITH MIMIC LOSS FOR ROBUST SPEECH RECOGNITION

Deblin Bagchi, Peter Plantinga, Adam Stiff & Eric Fosler-Lussier

The Ohio State University

Summary

For the task of speech feature denoising, local learning objectives are agnostic to phonetic structures helpful for speech recognition. We propose to add a global criterion to ensure denoised speech is useful for downstream tasks like ASR. This global criterion (mimic loss) is combined with the traditional local criterion to train the spectral mapper to produce denoised speech. This feature denoiser is independent of any particular acoustic model, and could be used as a pre-processor for any ASR system. This modularity is the strength of mimic loss.

Spectral Mapper

- After applying Short Term Fourier Transform (STFT) on noisy signal x, each spectral component x_m^k is augmented with deltas, double deltas and ten frames of context (designated $\tilde{x}_m^k = [x_{m\pm 5}^k]$).
- We define y_m to be the clean spectral slice at time m.
- We then use 2 layers of 2048 ReLU neurons to learn a mapping $f(\cdot)$ from noisy spectral slices $\tilde{x}_m$ to clean spectral features y_m using an MSE loss function, which we call *fidelity loss*.

$$\mathcal{L}_{\text{Fidelity}}(\tilde{x}_m, y_m) = \frac{1}{K}\sum_{k=1}^{K}(y_m^k - f(\tilde{x}_m)^k)^2$$

Spectral Classifier

- The spectral classifier is 6 layers of 1024 ReLU neurons, trained to classify a stacked clean spectral pattern $\tilde{y}_m$ as one of 1999 senone classes.
- We train the classifier using a cross entropy criterion. Once the classifier is trained, we freeze the weights.

Mimic Loss

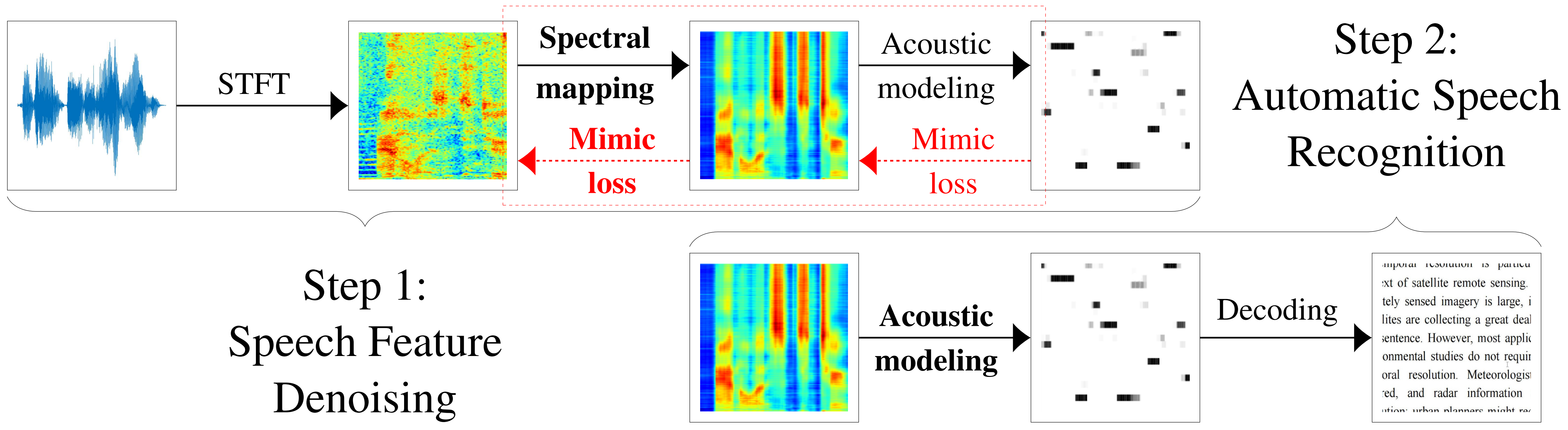
- We can define the *mimic loss* as the mean square difference between two D-dimensional representations generated by the spectral classifier $g(\cdot)$, one when evaluated on clean speech y_m, and one when evaluated on its paired denoised speech $f(\tilde{x}_m)$
- We experimented with two different representations for $g(\cdot)$: the posterior output of the senones after softmax normalization (*post-softmax*) and the layer outputs prior to the softmax normalization (*pre-softmax*).

$$\mathcal{L}_{\text{Mimic}}(\tilde{\tilde{x}}_m, \tilde{y}_m) = \frac{1}{D}\sum_{d=1}^{D}(g(\tilde{y}_m)^d - g(\tilde{f}(\tilde{x}_m))^d)^2$$

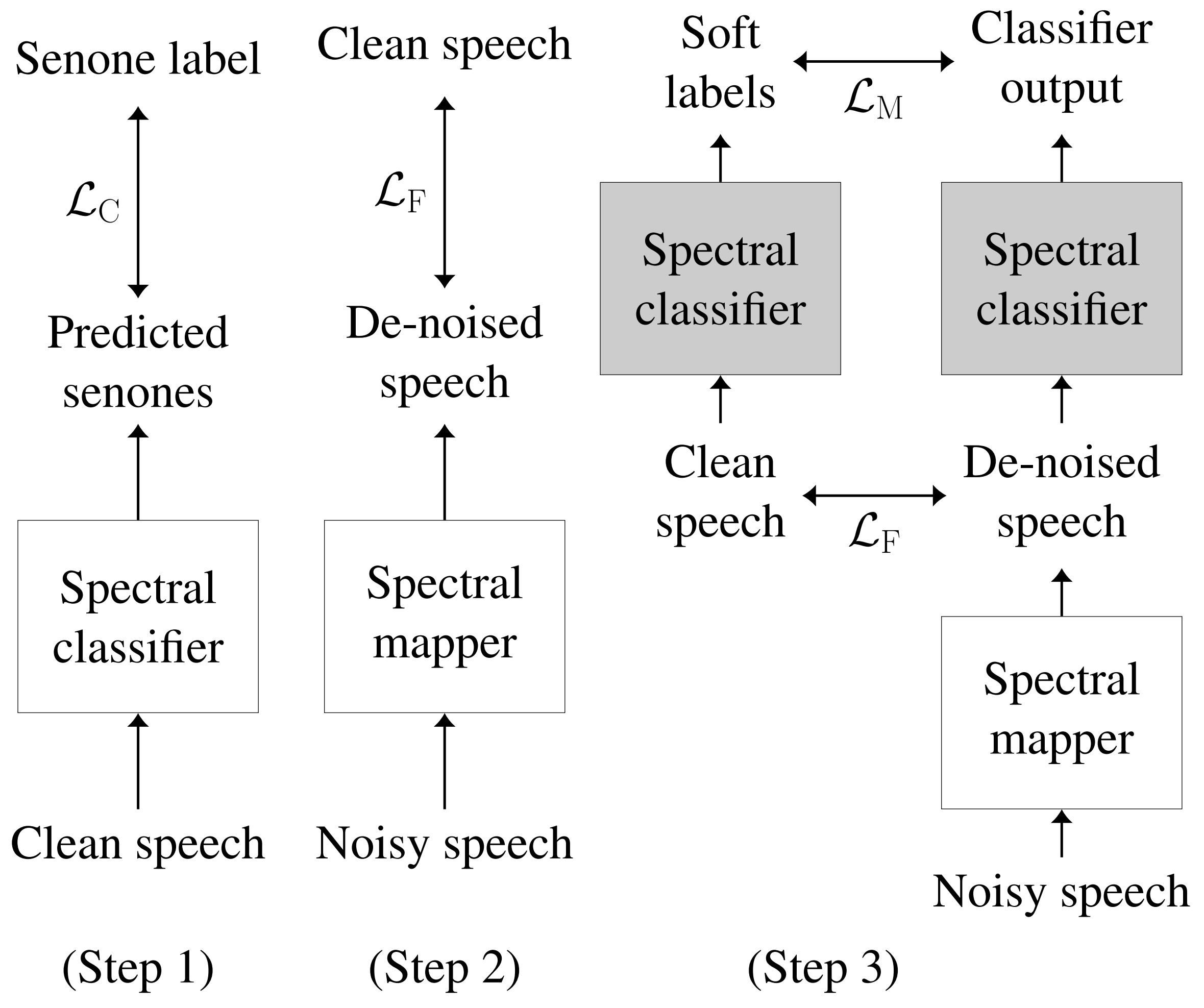
Joint Loss

- While training the spectral mapper, we found that mimic loss alone was not enough for the model to converge.
- We speculate that the task of predicting senones is too different from the task of predicting clean speech for the error signal to drive the output of the spectral mapper to actually look like speech features.
- Combining the fidelity and mimic losses into a joint loss allows the mapper to better imitate the behavior of the classifier under clean speech while keeping the denoised speech closer to clean speech.
- However, our approach should not be confused with joint training because our acoustic model (after pre-training) is frozen and only the spectral mapper is updated.

$$\mathcal{L}_{\text{Joint}} = \mathcal{L}_{\text{Fidelity}} + \alpha\mathcal{L}_{\text{Mimic}}$$



Spectral mapping system pipeline. Bold text indicates training a model.



Training with mimic loss. Gray indicates frozen weights.

Experiment and Results

- We evaluate the effectiveness of our proposed method on Track 2 of the CHiME-2 challenge by feeding denoised features to an off-the-shelf Kaldi recipe.

Spectral input to Kaldi	WER
Noisy features	17.3
Features denoised via fidelity loss	16.5
Features denoised via joint loss	
w/ post-softmax mimic loss	15.7
w/ pre-softmax mimic loss	**14.7**

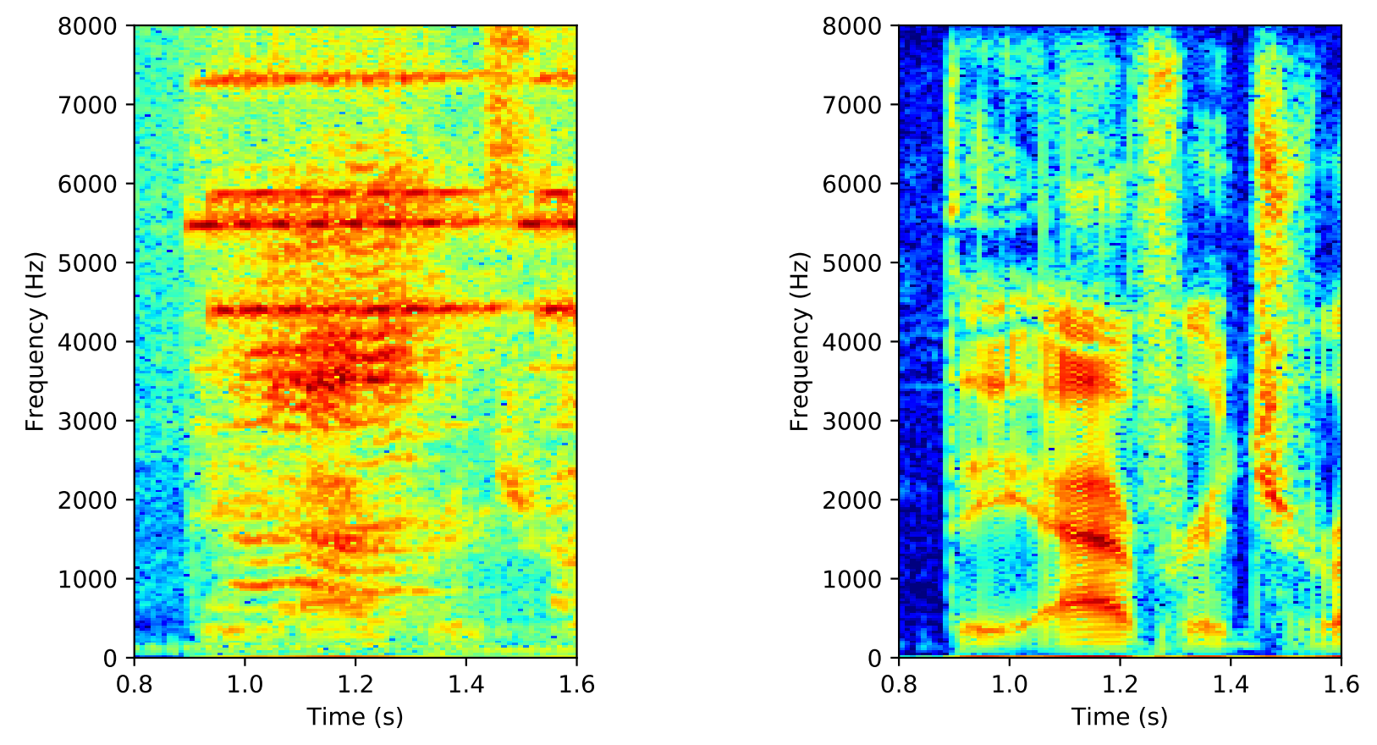
Experimental results on the CHiME2 test set

- Ongoing work suggests that a cross-entropy mimic loss on post-softmax targets performs similarly to MSE mimic loss on pre-softmax targets.
- Joint loss training diverges when the noisy speech recognizer is trained using hard targets rather than the soft targets of mimic loss which proves that this model benefits from learning to mimic the behavior of clean speech.
- For context, we show the performance of our system relative to other published results on this dataset. The better-performing models in this list use noise-robust features [1, 2, 3] as well as joint training of speech enhancement module and acoustic model [1, 3, 4] and more sophisticated models (like RNNs and CNNs) [2, 4].

Study	WER
Wang et.al [1]	10.6
Weninger et.al [2]	13.8
proposed approach	14.7
Narayanan-Wang [3]	15.4
Chen et. al [4]	16.0

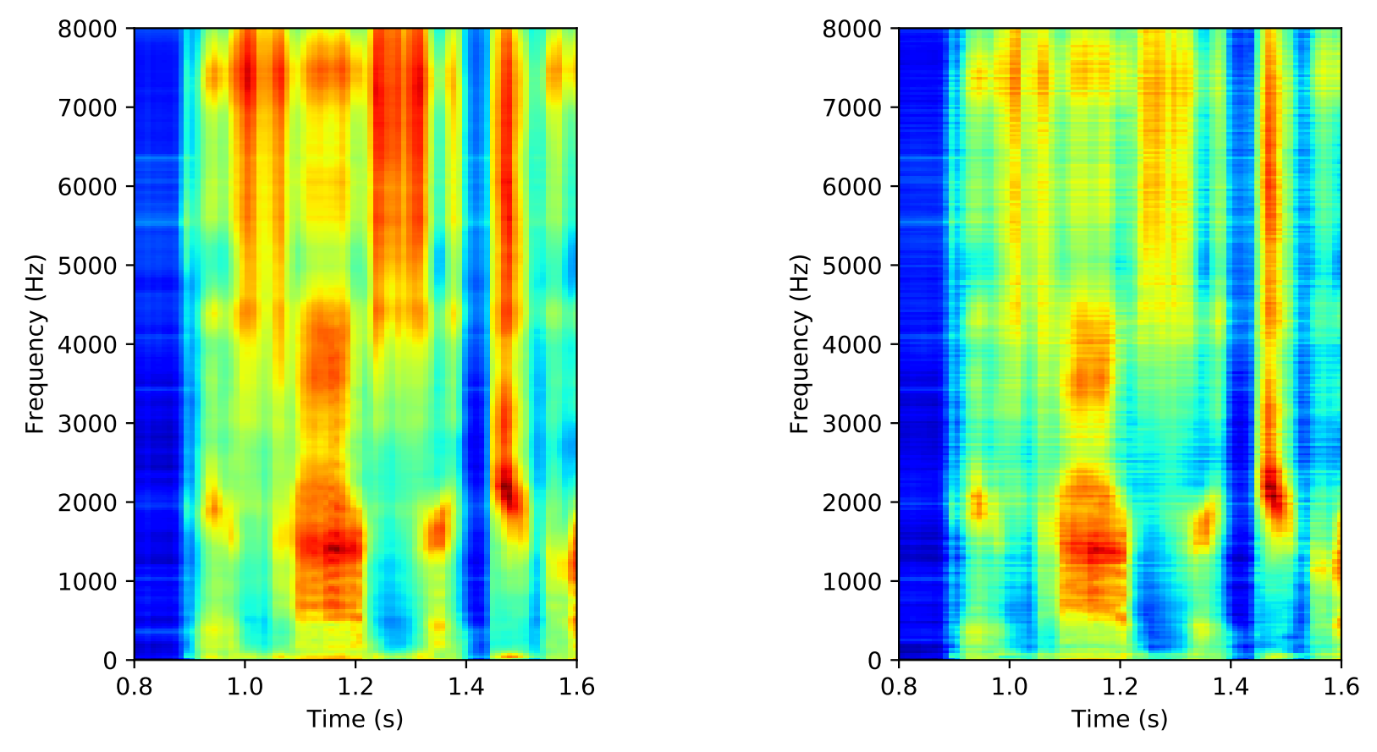
Performance comparison of proposed approach with other studies on the CHiME2 test set.

Spectrogram Comparison



(a) Noisy utterance **(b)** Clean utterance

Features for utterance 440c020f7 which begins with "The average rate . . . "



(c) Fidelity loss **(d)** Joint loss

Enhanced features for utterance 440c020f7. The Kaldi recipe incorrectly predicts the fidelity-loss-denoised features to have said "Disaster trade . . . ", but makes the correct prediction for the joint-loss-denoised features.

References

[1] Z.-Q. Wang and D. Wang, "A joint training framework for robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 796–806, 2016.

[2] 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.

[3] A. Narayanan and D. Wang, "Improving robustness of deep neural network acoustic models via speech separation and joint adaptive training," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 92–101, 2015.

[4] Z. Chen, S. Watanabe, H. Erdogan, and J. Hershey, "Integration of speech enhancement and recognition using long-short term memory recurrent neural network," in *Proc. Interspeech*, 2015.