1. Introduction
Speech enhancement aims to recover a clean speech signal from a noisy mixture signal.

Motivation
Gaussian noise modeling based on nonnegative matrix factorization (NMF) is common in semi-supervised speech enhancement, but it is limiting for certain types of noise.

Example noise signal recorded inside an underground subway:

Contribution
- Flexible alpha-stable noise model combined with a deep generative speech model for semi-supervised speech enhancement.
- Monte Carlo expectation-maximization (MCEM) algorithm.
- Outperforms the counterpart approach [1] based on Gaussian noise modeling with NMF variance parametrization.

2. Symmetric and positive alpha-stable distributions
Alpha-stable distributions are heavy-tailed distributions.

3. Supervised deep generative speech model
In the short-time Fourier transform domain, independently for all \((f, n) \in \mathbb{B} = \{0, \ldots, F-1\} \times \{0, \ldots, N-1\}\) we have [1, 2]:

\[
s_{fn} | h_n \sim N_c(0, \sigma_n^2, h_n)), \quad \text{where} \quad h_n \sim N(0,1).
\]

The parameters \(\theta_n\) of the neural network are learned from a dataset of clean speech signals. They are estimated by maximizing a lower bound of the log-likelihood, in the framework of variational autoencoders [3].

4. Unsupervised alpha-stable noise model
Marginal circularly symmetric alpha-stable noise model
Independently for all \((f, n) \in \mathbb{B}:

\[ b_{fn} \sim \mathcal{S}_0 S(\sigma_n, f). \]

where \(\sigma_n^2\) can be understood as the noise power spectral density (PSD).

Equivalent conditionally Gaussian noise model

\[ b_{fn} | \phi_{fn} \sim N_c(0, \phi_{fn} \sigma_n^2, f), \quad \text{where} \quad \phi_{fn} \sim \mathcal{P}(2 \cos(\pi a/4) / a). \]

\(\phi_{fn} \in \mathbb{R}\) is an impulse variable carrying uncertainty about the stationarity assumption of the marginal noise model.

5. Mixture model
The observed mixture signal is modeled as follows:

\[ x_{fn} = \sqrt{g_n} s_{fn} + b_{fn}, \]

where \(g_n \in \mathbb{R}_+\) represents a frame-dependent gain. We further consider the conditional independence of the speech and noise signals so that:

\[ x_{fn} | h_n, \phi_{fn} \sim N_c(0, g_n \sigma_n^2, h_n) + \phi_{fn} \sigma_n^2, f). \]

6. Inference
- Unsupervised noise model parameters to be estimated:
  \[ \theta_n = \left\{ g_n \in \mathbb{R}_+, \sigma_n^2 \in \mathbb{R}_{+, f} \right\} \]
- Observed variables: \(x = \{ x_{fn} \}_{(f,n) \in \mathbb{B}}\)
- Latent variables: \(z = \{ h_n, \phi_{fn} \}_{(f,n) \in \mathbb{B}}, (\mathbb{N} - 1)\)

MCEM algorithm
From an initialization \(\theta_0\) of the parameters, iterate:
- E-Step: \(Q(\theta_n; \theta_0) = \mathbb{E}_{z|x, \theta_n, \theta_0}[\ln p(x, z; \theta_n, \theta_0)]\).
  - Intractable expectation \(\rightarrow\) Markov chain Monte Carlo method.
- M-Step: \(\theta'_n = \arg\max_{\theta_n} Q(\theta_n; \theta_0)\).
  - Minorize-maximize approach leading to multiplicative update rules.

Posterior mean speech estimate with Wiener-like filtering
Let \(\hat{s}_{fn} = \sqrt{\mu} s_{fn}\) be the scaled speech STFT coefficients.

\[ \hat{s}_{fn} = \mathbb{E}_{p(s_{fn}|x, \theta_n, \theta_0)}[\delta_{fn}] = \mathbb{E}_{p(s_{fn}|x, \theta_n, \theta_0)} \left[ \frac{g_n \sigma_n^2, h_n + \phi_{fn} \sigma_n^2, f}{g_n \sigma_n^2, h_n + \phi_{fn} \sigma_n^2, f} \right] x_{fn}. \]

7. Reference method [1]
The speech model is the same, only the noise model differs. It is Gaussian with NMF-based variance parametrization:

\[ b_{fn} \sim N_c(0, (WH)_{fn}), \]

where both \(W \in \mathbb{R}^{K \times N}\) and \(H \in \mathbb{R}^{K' \times N}\) are estimated from the noisy mixture signal using an MCEM algorithm.

8. Experiments
- Training set (~4 hours): 462 speakers \times 10 sentences \times 3 seconds.
- Test set: 168 noisy mixtures (~3 seconds) at 0 dB SNR.
- Noise types: Domestic or office environments, nature, indoor public spaces, street, transportation.

References

- (SDR, SIR, SAR) measure (global quality, interferences, artifacts).
- (PESQ, STOI) measure (perceptual quality, intelligibility).