Log-Spectral Feature Reconstruction Based on an Occlusion Model for Noise Robust Speech Recognition

José A. González[a], Antonio M. Peinado[a], Angel M. Gómez[a], and Ning Ma[b]

[a] Dpt. Signal Theory, Telematics, and Communications, University of Granada, Spain
[b] Dpt. Computer Science, University of Sheffield, UK

Abstract

- The performance of ASR systems significantly degrades in the presence of noise.
- A new attempt to increase the robustness of ASR against additive noise is proposed here.
- **PROPOSAL**: Log-spectral feature reconstruction technique based on MMSE estimation and derived from a Log-Max observation model.
- Experimental evaluation:
  - Experiments were conducted on the Aurora2 and Aurora4 databases.
  - The proposed technique is compared with missing-data reconstruction.
  - Our proposal consistently outperforms missing-data reconstruction when using either binary or soft masks.

Occlusion Model

- Let $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{n}$ be the log-Mel filterbank energies corresponding to clean speech, noisy speech, and additive noise, respectively. The model that relates these variables is:

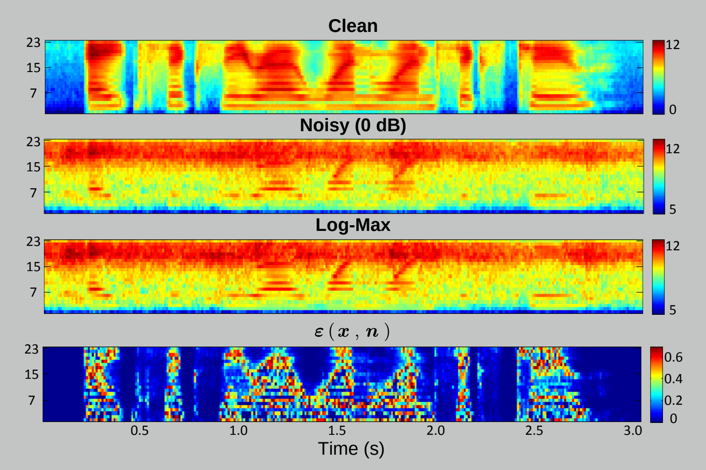
$$\boldsymbol{y} \approx \log(e^{\boldsymbol{x}} + e^{\boldsymbol{n}})$$

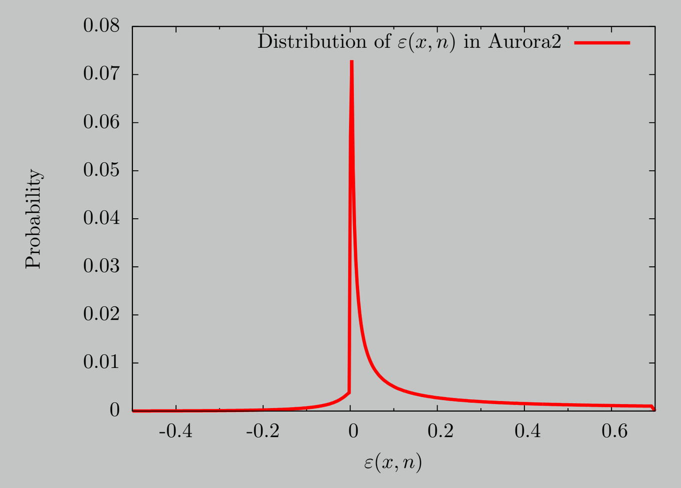
- This model can be rewritten as,

$$\boldsymbol{y} \approx \max(\boldsymbol{x}, \boldsymbol{n}) + \varepsilon(\boldsymbol{x}, \boldsymbol{n}) \qquad (1)$$

where

$$\varepsilon(\boldsymbol{x}, \boldsymbol{n}) = \log\left(\mathbf{1} + e^{\min(\boldsymbol{x},\boldsymbol{n}) - \max(\boldsymbol{x},\boldsymbol{n})}\right)$$





- Hence, $\varepsilon(\boldsymbol{x}, \boldsymbol{n})$ can be safely ignored from (1) and the resulting model is:

$$\boldsymbol{y} \approx \max(\boldsymbol{x}, \boldsymbol{n}) \qquad (2)$$

- Eq. (2) will be referred to as the **speech occlusion model** (also known as the Log-Max approximation in the literature).
- According to (2), the noisy feature vector can be rearranged into $\boldsymbol{y} = (\boldsymbol{y}_r, \boldsymbol{y}_u)$:
  - Reliable features ($\boldsymbol{x}_r \approx \boldsymbol{y}_r$), i.e. speech is not affected by noise.
  - Unreliable features ($-\infty \leq \boldsymbol{x}_u \leq \boldsymbol{y}_u$): speech is masked by noise.
- The occlusion model was first proposed by (Varga & Moore, 1990) to perform speech recognition with independent speech and noise HMMs.

Proposed Reconstruction Technique

- The **MMSE estimate** of the clean feature vector is given by

$$\hat{\boldsymbol{x}} = \mathbb{E}[\boldsymbol{x}|\boldsymbol{y}] = \int_{\boldsymbol{x}} \boldsymbol{x} p(\boldsymbol{x}|\boldsymbol{y}) d\boldsymbol{x} \qquad (3)$$

- Clean speech is modeled using a GMM λ_X.
- Noise distribution is $p(\boldsymbol{n}_t|\lambda_{N,t}) = \mathcal{N}_N(\boldsymbol{n}_t; \boldsymbol{\mu}_{N,t}, \Sigma_{N,t})$.
- Applying the above models, the MMSE estimate in (3) can be expressed as (time dependency is omitted),

$$\hat{\boldsymbol{x}} = \sum_{k=1}^{M} P(k|\boldsymbol{y}, \lambda_X, \lambda_N) \underbrace{\int_{\boldsymbol{x}} \boldsymbol{x} p(\boldsymbol{x}|\boldsymbol{y}, k, \lambda_X, \lambda_N) d\boldsymbol{x}}_{\mathbb{E}[\boldsymbol{x}|\boldsymbol{y}, k, \lambda_X, \lambda_N]} \qquad (4)$$

Posterior Computation

- Assuming independence among features, the corrupted speech likelihood is given by

$$p(y_i|k, \lambda_X, \lambda_N) = \iint p(y_i|x_i, n_i) p(x_i|k, \lambda_X) p(n_i|\lambda_N) dx_i dn_i \qquad (5)$$

- and the conditional likelihood is provided by the occlusion model in (2):

$$p(y_i|x_i, n_i) = \delta(y_i - \max(x_i, n_i)) = \delta(y_i - x_i)\mathbb{1}_{n_i \leq x_i} + \delta(y_i - n_i)\mathbb{1}_{x_i < n_i} \qquad (6)$$

- Then,

$$p(y_i|k, \lambda_X, \lambda_N) = p(y_i|k, \lambda_X) C(y_i|\lambda_N) + p(y_i|\lambda_N) C(y_i|k, \lambda_X) \qquad (7)$$

Partial Estimate Computation

- Proceeding in the same manner as before, the expectation in (4) can be obtained as,

$$\mathbb{E}[x_i|y_i, k, \lambda_X, \lambda_N] = w_i^k y_i + (1 - w_i^k)\tilde{\mu}_{X,i}^k \qquad (8)$$

- w_i^k is the speech presence probability:

$$w_i^k = \frac{p(y_i|k, \lambda_X)\, C(y_i|\lambda_N)}{p(y_i|k, \lambda_X, \lambda_N)}$$

- and $\tilde{\mu}_{X,i}^k$ is the mean of a right-truncated Gaussian pdf within the interval $(-\infty, y_i]$.

Comparison with Related Techniques

- Missing-data techniques (MDT) are also based on the occlusion model in (2).
- MDT assume that *a priori* knowledge about the feature reliability is provided by a missing-data mask $\boldsymbol{m}$.
- $\boldsymbol{m}$ can be either binary ($m_i = 1 \iff y_i$ reliable, $m_i = 0 \iff y_i$ unreliable) of soft ($m_i \in [0, 1]$).
- Then, $p(y_i|x_i, n_i)$ in (6) can be written now as,

$$p(y_i|x_i, n_i) = m_i \delta(y_i - x_i)\mathbb{1}_{n_i \leq x_i} + (1 - m_i)\delta(y_i - n_i)\mathbb{1}_{x_i < n_i} \qquad (9)$$

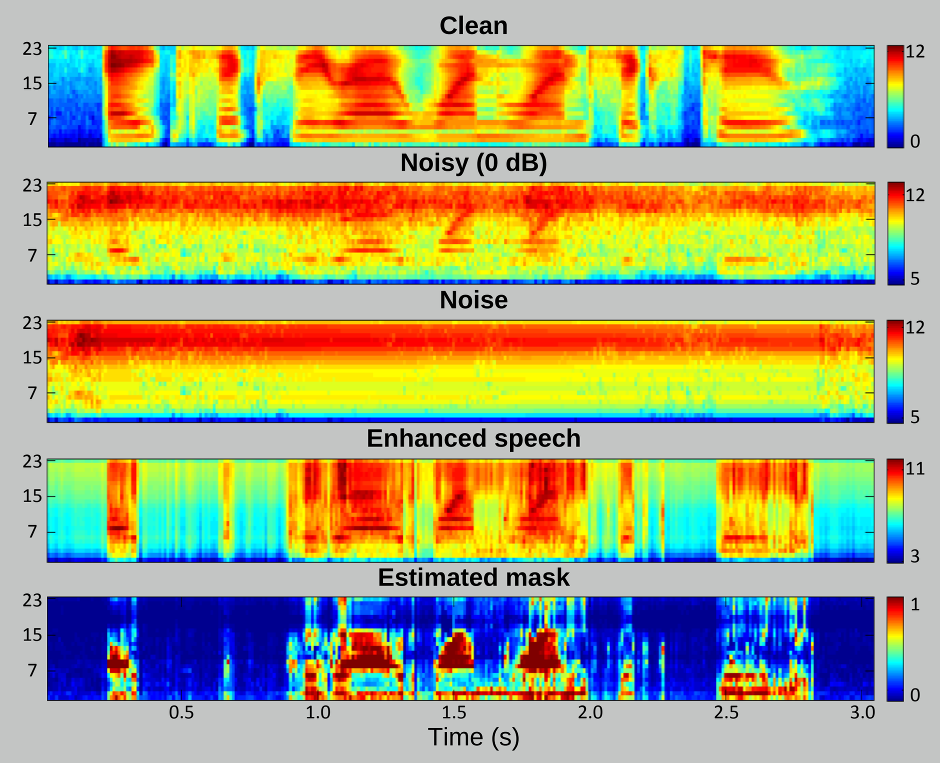
- Using (9) into (5), the terms required by the MMSE estimation in (4) can be computed as,

$$p(y_i|k, \lambda_X, \lambda_N) = m_i p(y_i|k, \lambda_X) C(y_i|\lambda_N) + (1 - m_i) p(y_i|\lambda_N) C(y_i|k, \lambda_X)$$
$$\mathbb{E}[x_i|y_i, k, \lambda_X, \lambda_N] = m_i y_i + (1 - m_i)\tilde{\mu}_{X,i}^k$$

Illustrative Example

- Aurora2 utterance (*eight six zero one one six two*) corrupted by subway noise at 0 dB.
- Estimated clean speech spectrogram is shown along with the estimated feature reliability mask.
- Feature reliability mask can be obtained as follows:

$$m_i = \sum_{k=1}^{M} P(k|\boldsymbol{y}, \lambda_X, \lambda_N)\, w_i^k \qquad (10)$$



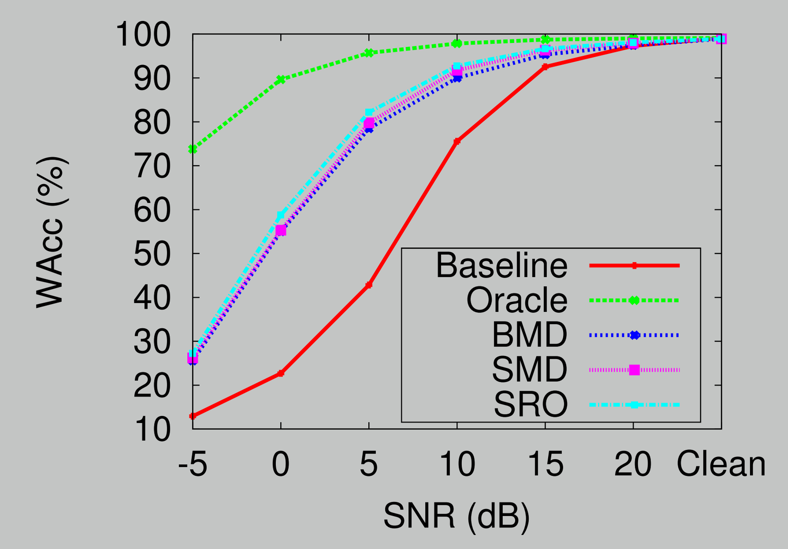
Experimental Results

Experimental Setup

- Speech features: 13 MFCCs + Δ + Δ^2 + CMN.
- Clean speech GMM with 256 components and diagonal covariances.
- Noise estimation: linear interpolation of the means for the N first and last frames ($N_{\text{Aurora2}} = 20$, $N_{\text{Aurora4}} = 40$). Σ_N fixed for the whole utterance.
- Mask computation: SNR estimates are thresholded (0 dB) to obtain the binary masks, whereas the soft masks are computed using (10).

Aurora2 Results

- Average results for Sets A, B, and C.
- Oracle: MDT with perfect knowledge of the feature reliability.
- Missing data techniques: BMD (binary masks) vs. SMD (soft masks).
- SRO (proposed technique) outperforms BMD and SMD.



Aurora4 Results

	T-01	T-02	T-03	T-04	T-05	T-06	T-07	Avg.	R.I.%
Baseline	87.69	75.30	53.24	53.15	46.80	56.36	45.38	**59.70**	-
Oracle	87.69	86.74	84.46	84.44	83.19	85.90	82.38	84.97	42.32
BMD	86.96	80.78	58.47	52.74	59.63	56.14	61.42	65.16	9.15
SMD	87.52	83.65	66.62	63.78	63.48	69.19	65.31	71.36	19.53
SRO	87.54	83.28	69.23	64.49	64.88	70.63	66.93	**72.43**	**21.31**

joseangl@ugr.es