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 \( x, y, \) and \( n \) be the log-Mel filterbank energies corresponding to clean speech, noisy speech, and additive noise, respectively. The model that relates these variables is:
\[
y = \log(\varepsilon - \sigma^2)
\]

This model can be rewritten as:
\[
y = \log(y(n) - x(n))
\]

where
\[
v(x, n) = \log(1 + e^{\delta(n) - \delta(x)})
\]

Hence, \( v(x, n) \) can be safely ignored from (1) and the resulting model is:
\[
y = \max(x(n))
\]

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 \( y - (y_r, y_f) \).

- Reliable features \( x_r \approx y_r \); i.e. speech is not affected by noise.
- Unreliable features \( x_f \approx y_f \); 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
  \[
  \hat{x} = \mathbb{E}[x|y] = \int x(x|y) dx
  \]

- Clean speech is modeled using a GMM \( \lambda_x \).
- Noise distribution is \( \mathbb{P}(n|x) = \lambda_{nx}(n|x_\mu, \Sigma_{nx}) \).
- Applying the above models, the MMSE estimate in (3) can be expressed as (time dependency is omitted),
  \[
  \hat{x} = \sum_{i=1}^{M} p(k|x, \lambda_x) x(k) + \sum_{i=1}^{M} p(k|x, \lambda_n) n(k)
  \]

**Posterior Computation**

- Assuming independence among features, the corrected speech likelihood is given by
  \[
  p(y(k|x, \lambda_x, \lambda_n)) - \int p(y(n|x, \lambda_n)|x(k), \lambda_n) d\lambda_n
  \]

- and the conditional likelihood is provided by the occlusion model in (2):
  \[
  p(y(k|x, n)) = \delta(y(k) - \max(n, \lambda(n))) - \delta(y(k) - \lambda(k) \lambda(n)_\Sigma)
  \]

- Then,
  \[
  p(y(k|x, \lambda_x, \lambda_n)) - p(y(k|x, \lambda_n) \Omega(y(k)|x(k), \lambda_n)
  \]

**Partial Estimate Computation**

- Proceeding in the same manner as before, the expectation in (4) can be obtained as:
  \[
  \mathbb{E}[x(k)] = \delta(y(k) - \max(n, \lambda(n))) + \delta(y(k) - \lambda(k) \lambda(n)_\Sigma)
  \]

- \( \delta_n \) is the mean of a right-truncated Gaussian pdf within the interval \( [-\infty, y_r] \).

**Comparison with Related Techniques**

- Missing-data techniques (MDT) are also based on the occlusion model in (2).
- MDT assume that a priori knowledge of the feature reliability is provided by a missing-data mask \( m \).
- \( m \) can be either binary (\( m = 1 \) reliable, \( m = 0 \) unreliable) or soft \( m \in [0, 1] \).
- Then, \( p(y, n) \) in (6) can be written as:
  \[
  p(y, n) = p(y|x, \lambda_x) \Omega(n|x, \lambda_n)
  \]

- Using (9) into (5), the terms required by the MMSE estimate in (4) can be computed as:
  \[
  p(y(k|x, \lambda_x, \lambda_n)) - p(y(k|x, \lambda_n) \Omega(n|x, \lambda_n)
  \]

**Illustrative Example**

- Aurora2 utterance (eight six zero one one six two) corrupted by subway noise at 0 dB.
- Estimated clean speech spectrum is shown along with the estimated feature reliability mask.
- Feature reliability mask can be obtained as follows:
  \[
  m = \sum_{i=1}^{M} p(k|x, \lambda_x, \lambda_n) w_i^2
  \]

**Experimental Results**

- **Experimental Setup:**
  - Speech features: 13 MFCCs + \( \Delta \Delta + \) CMN.
  - Clean speech GMM with 256 components and diagonal covariances.
  - Noise estimation: linear interpolation of the means for the N first and last frames \( \lambda_{n_{0\text{ord}}} = 0, \lambda_{n_{2\text{ord}}} = 45 \). \( \lambda_{n_{0\text{ord}}} \) 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**

- Table below shows results obtained for each condition.
- The best results are highlighted in bold.
- SRO outperforms BMD and SMD in almost all conditions.