Analysis-by-synthesis feature estimation for robust automatic speech recognition using spectral masks

Michael I Mandel & Arun Narayanan · Computer Science & Engineering · {mandelm,narayaar}@cse.osu.edu

PROBLEM

• Automatic speech recognition (ASR) is easily disrupted by background noise
• Spectral masking suppresses noise well for listeners, but leads to distorted ASR features that are not recognized accurately
• Can we estimate ASR features of the clean speech from a masked noisy mixture?

SOLUTION

• Use full recognizer trained on clean speech as prior model in reconstructing clean ASR features
• Find latent clean speech features that minimize – Itakura-Saito divergence between synthesized spectral envelope and noisy observation, weighted by mask
• Negative log likelihood under lattice from full large vocabulary continuous speech recognizer
• Optimize using gradient descent

BENEFITS

• Uses unmodified full recognizer as prior
• Flexible framework, easy to add new terms
• Accommodates masks in different domains
• Reduces word error rates, even for oracle masks
• Reduces distance of estimates to clean speech

DIFFERENTIABLE ANALYSIS-BY-SYNTHESIS PIPELINE

- Compressed Aud Spectrum
- Surround: Warped Frequency: Power Spectrum
- MFCC: x
- Spectral masking
- Weighted comparison (c)
- Un warped Frequency: Power Spectrum
- MFCC-39
- Mean-Vari
- Normalized MFCC-39
- Add terms to cost function, e.g., penalize energy
- Apply to deep neural net acoustic models
- Fully joint recognition and separation
- Add terms to cost function, e.g., penalize energy in resynthesis that is louder than mixture
- Test on lower SNRs (e.g., CHiME2)

MATHEMATICAL DETAILS

For a given MFCC matrix, we synthesize a power spectrum matrix, \( \hat{S}_{\text{act}}(x) \), and compare it to the noisy observed power spectrum matrix, \( S_{\text{act}} \), in regions selected by a mask, \( M_{\text{act}} \), creating the Itakura-Saito cost:

\[
L_I(x; M_{\text{act}}) = D_W \left( S_{\text{act}} \parallel \hat{S}_{\text{act}}(x) \right) = \sum_{\omega,t} W_{\omega t} \left( \frac{S_{\text{act}}(\omega t)}{\hat{S}_{\text{act}}(\omega t)} - \log \hat{S}_{\text{act}}(\omega t) \right) - 1 \quad (1)
\]

This cost is then combined with the log likelihood of the ASR features computed from \( x, y(x) \), under the large vocabulary ASR system, \( L_H(x) \), to give the combined cost function (with tradeoff \( \alpha \))

\[
L(x; M) = (1 - \alpha)L_I(x; M) + \alpha L_H(y(x)) \quad (2)
\]

- Minimize using quasi-Newton BFGS (unconstrained, nonlinear, nonconvex optimization)
- Numerical gradient of \( L_I \) computed for each frame separately
- Closed form gradient of \( L_H \) computed using forward-backward algorithm

RESULTS

<table>
<thead>
<tr>
<th>Mask</th>
<th>Lattice</th>
<th>WER</th>
<th>Direct</th>
<th>( \Delta ) Abs</th>
<th>IS div to clean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean</td>
<td>Oracle</td>
<td>14.38</td>
<td>12.99</td>
<td>9.54</td>
<td>273007</td>
</tr>
<tr>
<td>Est. Clean</td>
<td>16.18</td>
<td>15.31</td>
<td>13.53</td>
<td>115479</td>
<td>–1414</td>
</tr>
<tr>
<td>Est.</td>
<td>16.18</td>
<td>15.31</td>
<td>13.53</td>
<td>274679</td>
<td>–1273</td>
</tr>
<tr>
<td>Noisy</td>
<td>30.94</td>
<td>27230</td>
<td>–549</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Add terms to cost function, e.g., penalize energy in resynthesis that is louder than mixture

EXAMPE: RESTAURANT NOISE, 9 DB SNR, ORACLE MASK

- Transcription it01: “He said such products would be marketed by other companies with experience him at this month.”
- Transcription it04: “He said such products would be marketed by other companies with experience in that business.”

REFERENCES


APDX: HMM GRAD. WRT OBS.

For observation \( y_k \), \( y_{k-1} \) from \( k = 1 \) to \( N \), the HMM log likelihood is

\[
L_H(x | y_1 \ldots y_N) = \log \sum_{i=1}^{N} a_{i1} b_{i1} \pi_1 y_1 \ldots y_N \quad (9)
\]

Gradient of the log likelihood with respect to a particular \( y_k \) for GMM emissions, \( b_{ik} \), in

\[
\frac{\partial L_H(x | y_1 \ldots y_N)}{\partial y_k} = \sum_{i=1}^{N} a_{ik} b_{ik} \frac{1}{y_k} \pi_1 y_1 \ldots y_N - \sum_{i=1}^{N} a_{ik} b_{ik} \frac{1}{y_k} \pi_1 y_1 \ldots y_{k-1} \sum_{i=1}^{N} a_{ik} b_{ik} \sum_{i=1}^{N} a_{ik} b_{ik} y_{k+1} \ldots y_{N} \quad (10)
\]