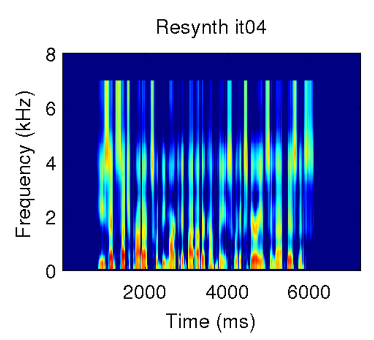


Analysis-by-synthesis feature estimation for robust automatic speech recognition using spectral masks

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

The Ohio State University

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?

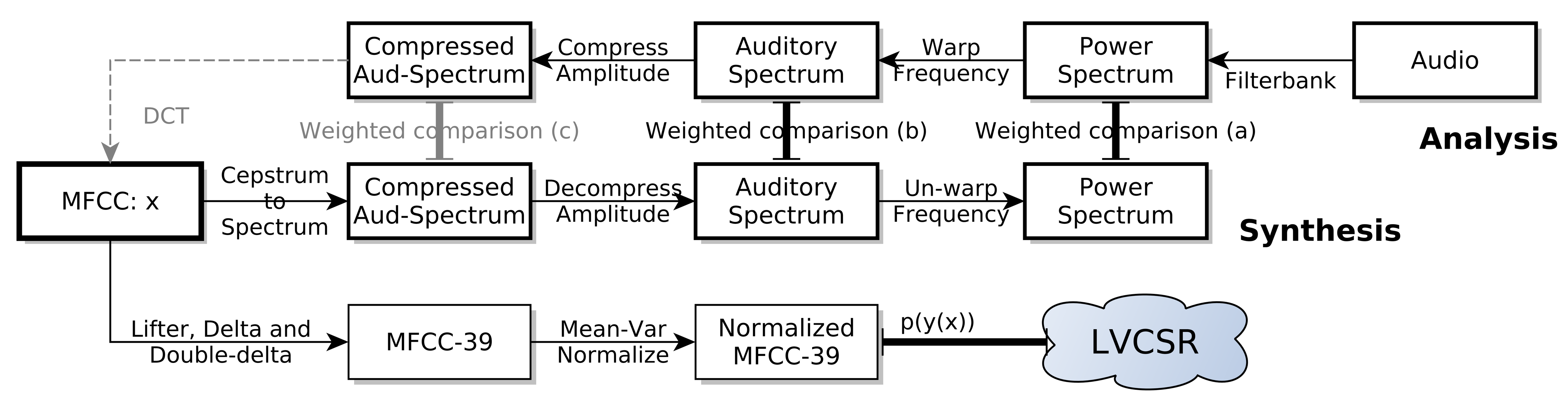
Solution

- 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



Mathematical details

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

$$\mathcal{L}_I(\mathbf{x}; M_{\omega t}) = D_W\left(S_{\omega t} \| \tilde{S}_{\omega t}(\mathbf{x})\right) = \sum_{\omega,t} W_{\omega t}\left(\frac{S_{\omega t}}{\tilde{S}_{\omega t}(\mathbf{x})} - \log\frac{S_{\omega t}}{\tilde{S}_{\omega t}(\mathbf{x})} - 1\right) \quad (1)$$

This cost is then combined with the log likelihood of the ASR features computed from $\mathbf{x}$, $y(\mathbf{x})$, under the large vocabulary ASR system, $\mathcal{L}_H(\mathbf{x})$, to give the combined cost function (with tradeoff α)

$$\mathcal{L}(\mathbf{x}; M) = (1-\alpha)\mathcal{L}_I(\mathbf{x}; M) + \alpha\mathcal{L}_H(y(\mathbf{x})) \quad (2)$$

- Minimize using quasi-Newton BFGS (unconstrained, nonlinear, nonconvex optimization)
- Numerical gradient of $\mathcal{L}_I$ computed for each frame separately
- Closed form gradient of $\mathcal{L}_H$ computed using forward-backward algorithm

Results

		WER		IS div to clean	
Mask	Lattice	Direct	AbS	Direct	Δ AbS
Clean		9.54			
Oracle	Clean	14.38	**12.99**	273006	−549
Est.	Clean	16.18	**13.53**	276497	−1419
Oracle	Est.	14.38	**13.62**	273006	−500
Est.	Est.	16.18	**15.31**	276497	−1273
Noisy		30.94		272301	

Word error rates and average Itakura-Saito divergence to the clean speech (IS div to clean) for direct-masking (Direct) and analysis-by-synthesis (AbS) feature estimation on AURORA4 task (996 test utterances, SNR 5–15 dB). Averaged over all 6 noise conditions. Recognizer trained on clean speech. Oracle mask is ideal binary mask, estimated (Est.) mask from Narayanan and Wang (2014). Estimated lattice from reconstructed features, cheating clean lattice from clean features.

Future Directions

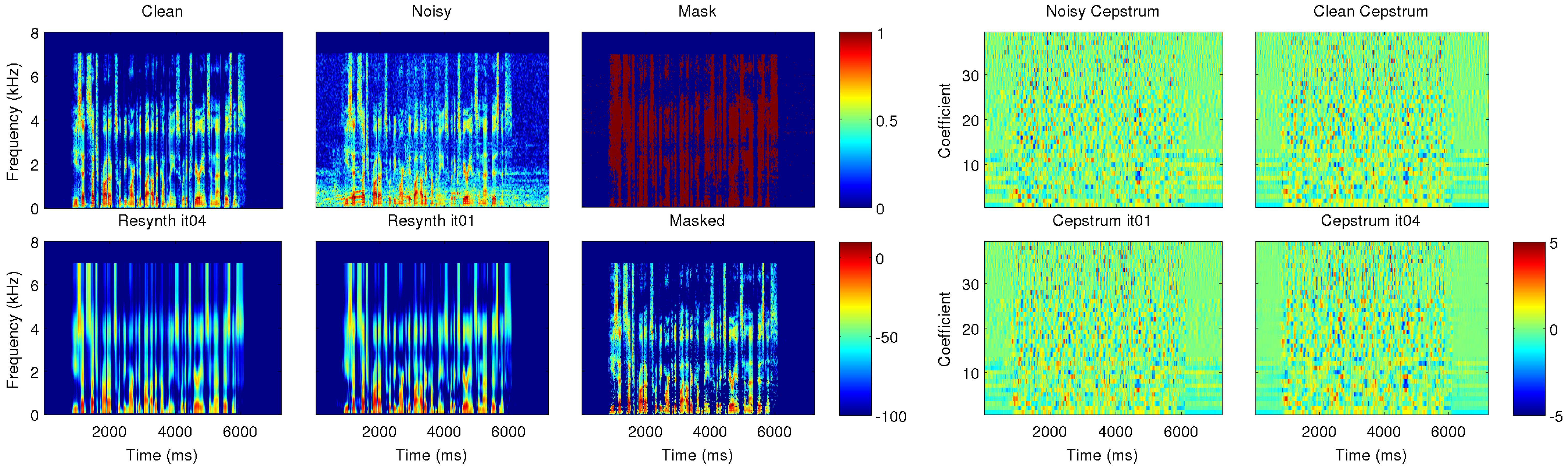
- Apply to deep neural net acoustic models
- Model speech excitation
- Fully joint recognition and separation
- Add terms to cost function, e.g., penalize energy in resynthesis that is louder than mixture
- Test on lower SNRs (e.g., CHiME2)

References

Narayanan, A. and D. Wang (2014). "Investigation of speech separation as a front-end for noise robust speech recognition". In: IEEE/ACM Tr. ASLP 22, pp. 826–835.

Example: 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."



Apdx: HMM grad. wrt obs.

For observations $y_{1:T}$ derived from $\mathbf{x}_{1:T}$, the HMM log-likelihood is

$$\mathcal{L}_H(y_{1:T}) = \log\sum_{i=1}^{I}\alpha_{it}\beta_{it} \quad \forall t \in \{1, \ldots, T\} \quad (3)$$

Gradient of the log likelihood with respect to a particular y_t for GMM emissions, b_{it}, is

$$\frac{\partial\mathcal{L}_H(y_{1:T})}{\partial y_t} = \frac{\partial}{\partial y_t}\log\sum_{i=1}^{I}\alpha_{it}\beta_{it} = p(y_{1:T})^{-1}\sum_{i=1}^{I}\frac{\alpha_{it}\beta_{it}}{b_{it}}\frac{\partial b_{it}}{\partial y_t} \quad (4)$$

$$\frac{\partial b_{it}}{\partial y_t} = \sum_{k=1}^{K}\pi_{ik}\mathcal{N}(y_t \mid \mu_{ik}, \Sigma_{ik})\Sigma_{ik}^{-1}(\mu_{ik} - y_t) \quad (5)$$