ASR-Driven Binary Mask Estimation using Spectral Priors

William Hartmann and Eric Fosler-Lussier (hartmann.59@osu.edu ; fosler@cse.ohio-state.edu)

Department of Computer Science and Engineering, The Ohio State University

Introduction

- One method of handling noise in a speech signal is to remove the noise prior to computing standard ASR features.
- The Ideal Binary Mask (IBM) has been proposed as a goal for speech separation (Wang, 2005).
- Typical binary mask estimation techniques focus on low-level features.
- We propose an alternative approach to mask estimation that forces ASR-driven, high-level features.
- We show improvements in both word error rate (WER) and signal to noise ratio (SNR).

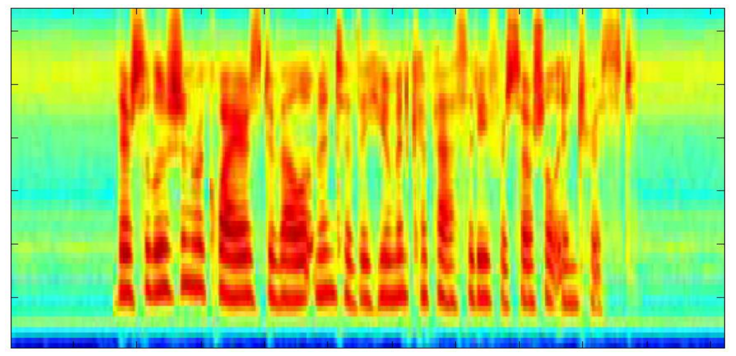
The Ideal Binary Mask

- The IBM is calculated on a spectro-temporal representation of the signal.
- Since we assume the noise is additive, we can measure the energy contribution of both the noise and speech at each time-frequency (T-F) unit.
- All noise-dominant T-F units are masked and speech-dominant T-F units are left unmasked.
- More formally, we can define the IBM as

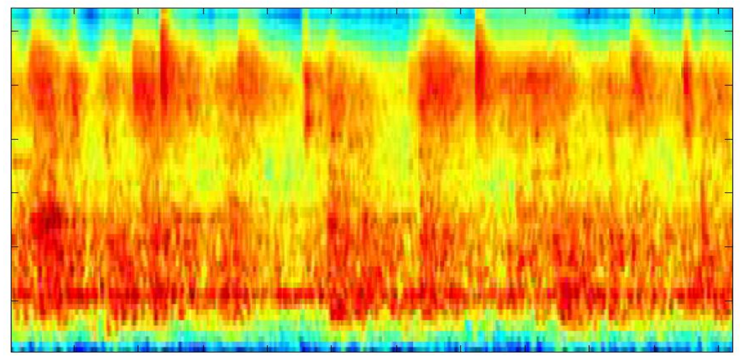
$$M(f,t) = \begin{cases} 1 & \frac{|S(f,t)|^2}{|N(f,t)|^2} > \theta \\ 0 & \text{otherwise} \end{cases}$$

- Given a priori knowledge of the speech signal and interfering noise, we can calculate the IBM.
- Below we show an example with Factory noise mixed at an SNR of 5 dB

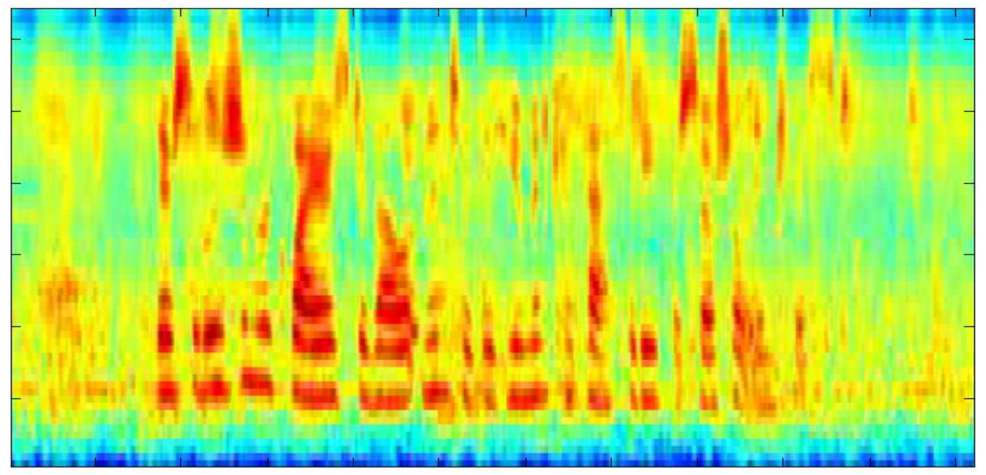
Clean Speech



Factory Noise



Speech mixed with Noise



Ideal Binary Mask

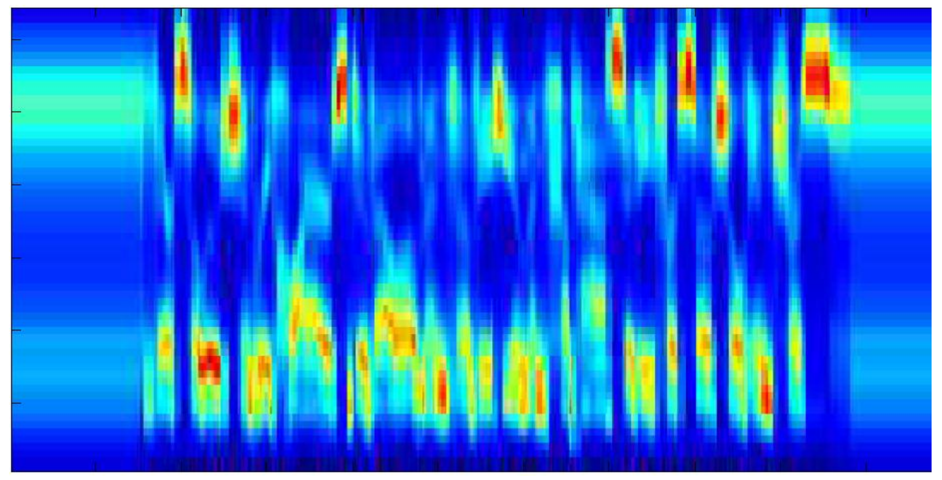
ASR-Driven Binary Mask

- In order to force the use of higher level information, we change the masking criterion.
- Instead of focusing on the interfering noise, we focus on the expected spectral characteristics of the speech.
- We use the alignments from an HMM to label the subphonetic units in a speech signal.
- For each subphonetic unit we create a model of the average distribution of energy. (e.g. k-ah+t[2], sil[1], t+s[3])
- We can create an oracle mask by comparing the model for the true subphonetic units to a background prior. See Figure below.
- During estimation, we do not know the true subphonetic unit.
  - We can collect a set of possible models by creating a 100-best list from the original mixed signal.
  - The model which best matches a conservative mask is then used.
- Explicitly, we define our criterion as

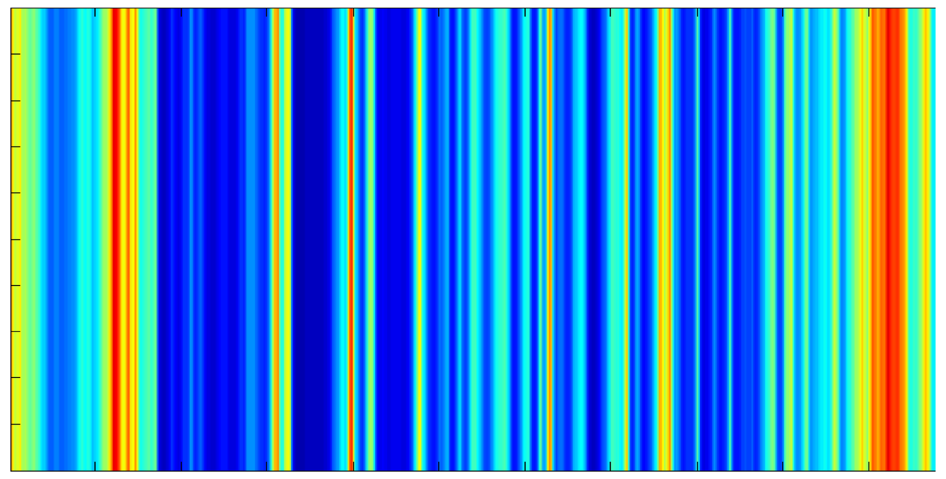
$$M(f,t) = \begin{cases} 1 & \alpha_{f,s_t} > \beta_f r_t^{\gamma} \\ 0 & \text{otherwise} \end{cases} \quad \text{where} \quad r_t = \frac{\text{average frame energy}}{\text{frame energy in frame } t}$$

ASR-Driven Mask Example

Spectral Prior



Weighted Background Prior



Oracle ASR-Driven Binary Mask

Experimental Results

- Estimated masks are generated from the candidate models collected from N-best lists.
  - 1-Best refers to the single best output from the ASR system.
  - 100-Best refers to the top 100 hypotheses from the ASR system.
- Oracle masks are generated by selecting the best model from a set of candidate models.
  - They represent a lower bound for word error rate.
- Results are obtained using an HMM-based recognizer built with HTK (Young et al., 2002).
- Features are mean-subtracted, variance normalized PLPs.

System	Car	Babble	Restaurant	Street	Airport	Train	Avg
Baseline	27.3	34.3	36.7	39.3	35.0	42.0	35.8
1-Best Estimate	25.2	32.5	35.5	37.7	33.4	39.7	34.0
100-Best Estimate	23.9	30.7	34.3	35.4	33.8	36.6	32.5
Oracle Masks							
Ideal Binary Mask	17.6	15.8	15.4	19.5	16.2	19.6	17.4
Clean Speech Oracle	19.0	20.1	24.1	20.5	22.6	21.6	21.3
100-Best Oracle	20.5	25.6	28.1	29.9	27.3	32.1	27.3

Word error rate on the Aurora4 corpus. Lower numbers are better.

- We also report speech enhancement results in terms of SNR improvement.
- Results are similar to a standard PSD-based enhancement system.
- Our system uses no knowledge regarding the underlying interference.

System	Car	Babble	Restaurant	Street	Airport	Train	Avg
Hendriks et al.	8.3	2.8	2.3	6.7	2.4	5.7	4.7
100-Best Estimate	10.9	3.1	2.3	7.1	2.8	6.0	5.4

SNR improvement on Aurora4. Comparison system is a standard PSD-based speech enhancement algorithm (Hendriks et al., 2010). Greater numbers are better.

Conclusions

- A binary mask, similar to the IBM, defined on the underlying linguistic content of the signal can produce significant WER improvements over an unenhanced baseline while ignoring the interfering noise.
- SNR improvements using the ASR-Driven binary mask are comparable to a standard speech enhancement technique.
- Noisy ASR results can drive the speech enhancement process.
- Future work will seek to improve the subphonetic model selection from the candidate models.

References

- Hendriks et al., MMSE-based noise PSD tracking with low complexity, in Proceedings of ICASSP, 2010, pp 4266-4269.
- S. Young et al., *The HTK Book*. Cambridge University Publishing Department, 2002. Available: http://htk.eng.cam.ac.uk/
- D.L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech separation by humans and machines*, P. Divenyi, Ed., pp. 181-197. Kluwer Academic, Norwell MA, 2005.

We gratefully acknowledge William Hartmann's support under NSF CAREER grant IIS-0643901. The opinions and conclusions expressed in this work are those of the authors and not of any funding agency.