E-Lobes

Mask-assisted speech enhancement for binaural hearing aids

ELOBES2019 workshop – 12 January 2019

Mike Brookes, Leo Lightburn, Alastair Moore,
Patrick Naylor & Wei Xue

Outline

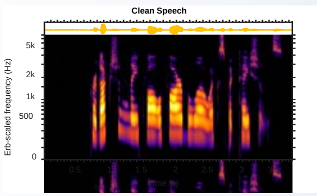
E-Lobes

- Motivation: Ideal Binary Mask (IBM)
  - Intelligibility model for IBM-masked speech
  - STOI-optimal binary mask and its estimation
- Mask-assisted MMSE enhancement
  - Single-channel performance
- Binaural Enhancement
  - Alternatives for Metric reference signals
  - Bilateral versus Binaural beamforming
  - Effect of an improved mask
- Summary

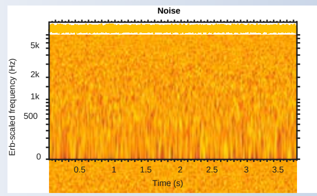
“Ideal” Binary Masks (IBM)

E-Lobes

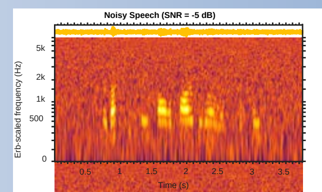
- Additive noise



+

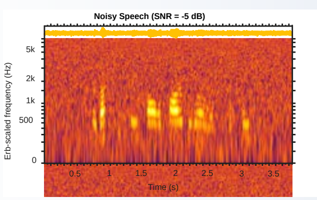


=

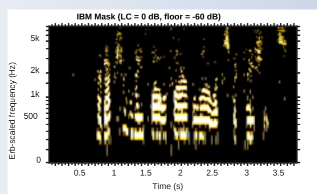


SNR = –5 dB
White Noise

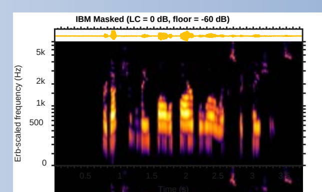
- Apply Binary Mask
  - Keep only time-frequency cells with local SNR > “local criterion” threshold (LC)



×



=



LC = 0 dB

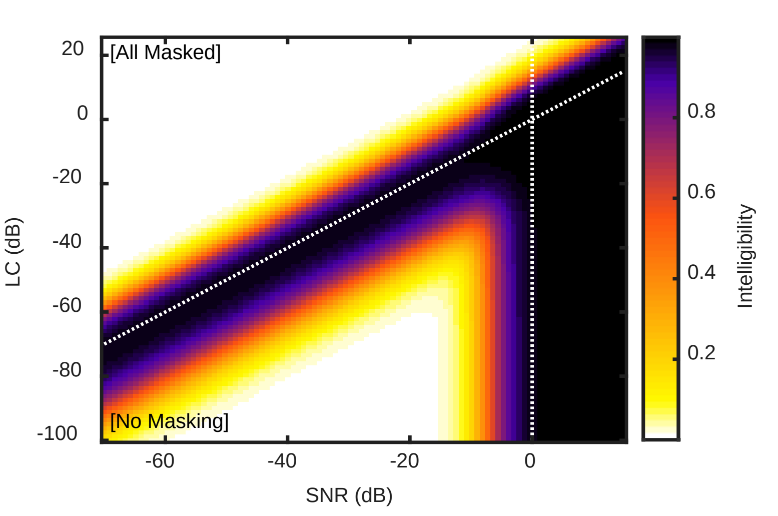
- An “oracle” mask has access to both the clean speech and the noise
  - In practice, the mask must be estimated from the noisy speech alone

IBM-Masked Speech Intelligibility

E-Lobes

∃ two independent sources of information: [Kjems et al 2010]

1. Noisy speech signal
   Distorted by the mask
2. Noise-vocoded signal
   Noise modulated by the mask



[from Kjems et al 2010]

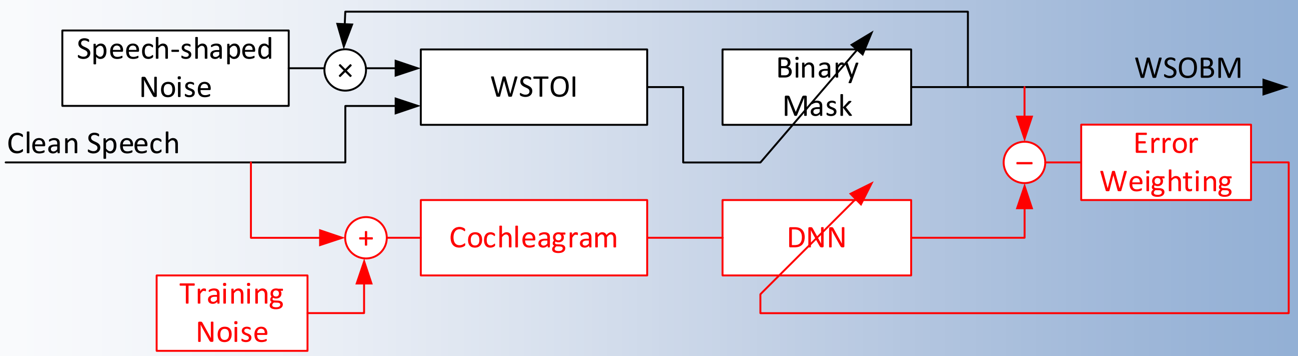
- Component 1 is intelligible for $SNR >\approx -5$ dB provided mask is not too sparse ($SNR > LC - 5$ dB)
  - vertical bar on figure
- Component 2 is intelligible if (a) high speech power → mask on ($SNR > LC - 5$ dB) and (b) low speech power → mask off ($SNR < LC + 20$ dB)
  - diagonal bar on figure

(1) The benefit of binary masking comes entirely from component 2

(2) The mask should reflect clean speech energy (not the local SNR)

STOI-optimal Binary Mask

E-Lobes

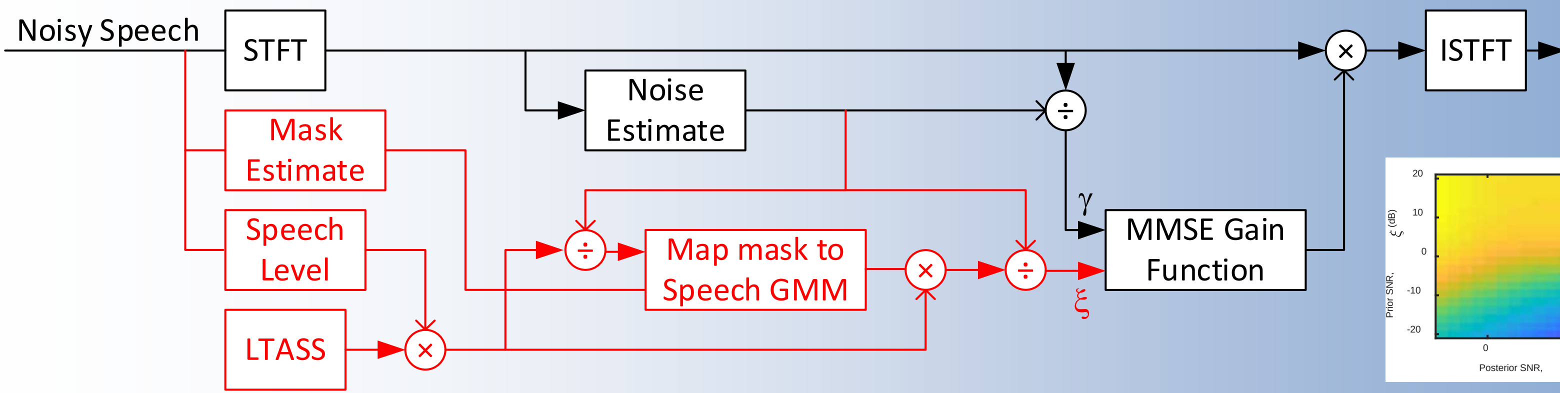


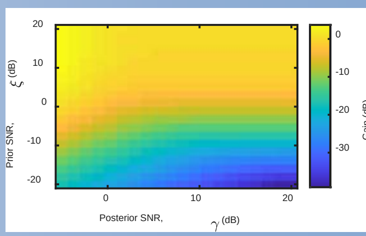
[Lightburn et al, 2015, 2016]

- The STOI-optimal binary mask (SOBM) maximizes the STOI of masked speech-shaped noise (SSN)
  - Depends only on the clean speech
  - WSTOI weights time-frames by estimated speech information
- Train DNN to estimate the mask from noisy speech
  - Trained on a range of noises at a range of SNRs
  - Error weighting: (a) freq band importance, (b) WSTOI sensitivity
  - DNN output $\in [0, 1]$ corresponds to probability that mask = 1

Mask-assisted Enhancement

E-Lobes





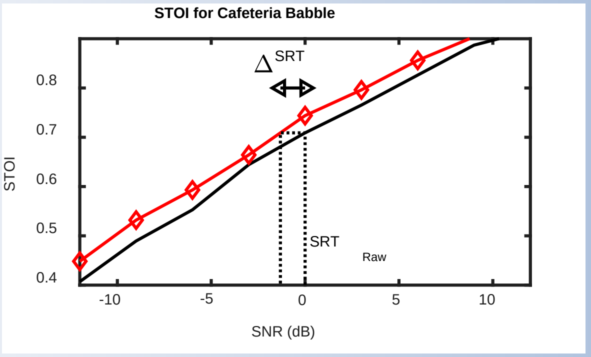
- LogMMSE enhancer assumes zero-mean complex Gaussian speech and noise STFT coefficient distributions
  - Gain function depends on posterior SNR, γ, and prior SNR, ξ
- Map mask to Gaussian Mixture Model (GMM) distribution for speech power
  - Mapping depends on frequency band and estimated SNR
  - Denormalize by estimated speech level in the frequency band
  - Divide by estimated noise power to get GMM for prior SNR, ξ

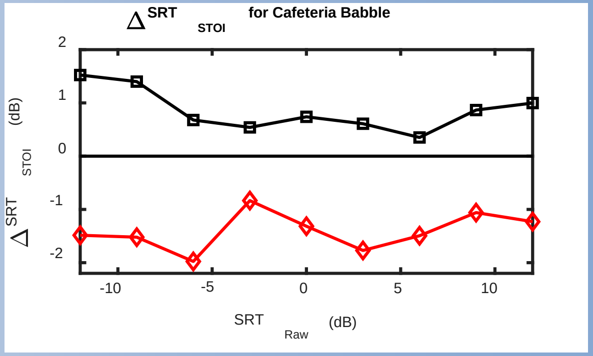
[Gonzalez et al, 2013; Ephraim et al, 1985]

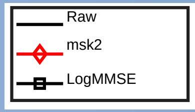
Single-channel Enhancement

E-Lobes

- Raw speech has acceptable intelligibility @ SNR=SRT_{Raw}
- Enhanced speech has the same intelligibility @ $SRT_{Raw}+\Delta SRT$

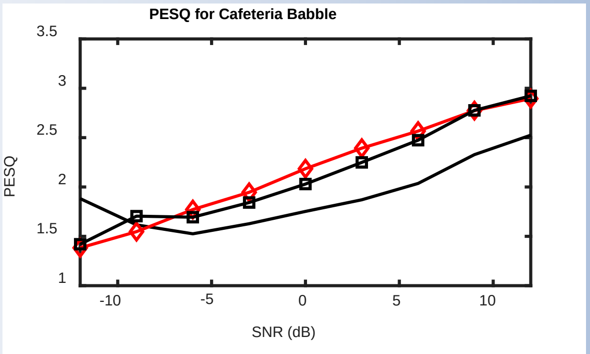


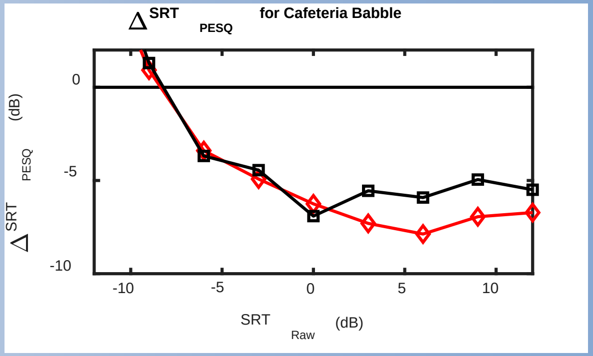




- Can regard $-\Delta SRT$ as increased tolerance to noise
- Mask-assisted enhanced has ΔSRT of –1.5 dB
- In contrast, LogMMSE enhancer has ΔSRT of +1 dB

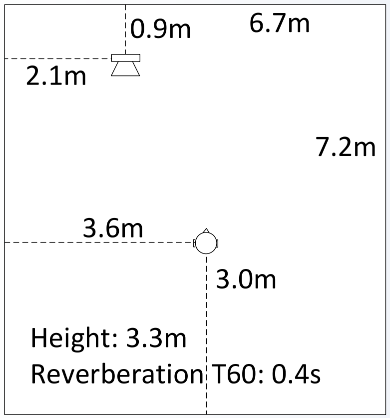
- PESQ tolerance to noise improves by >5 dB for both enhancers at SNR_{Raw} > –5 dB
  - Note: PESQ unreliable at low SNRs.

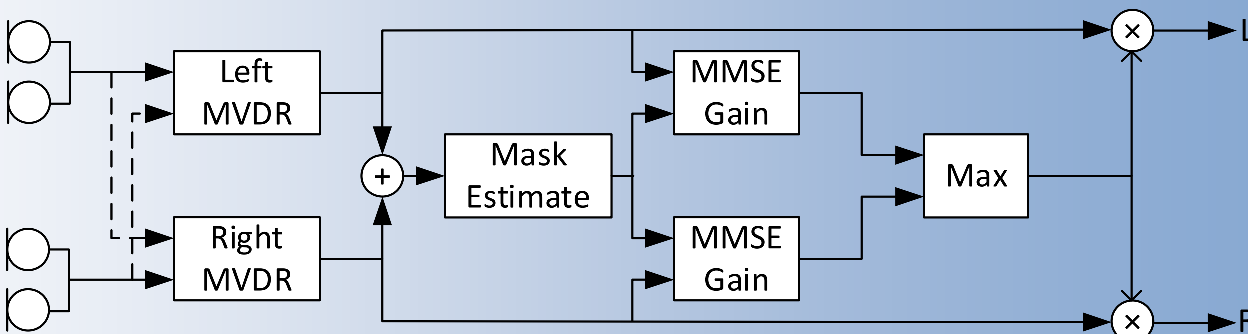




Binaural Enhancement

E-Lobes





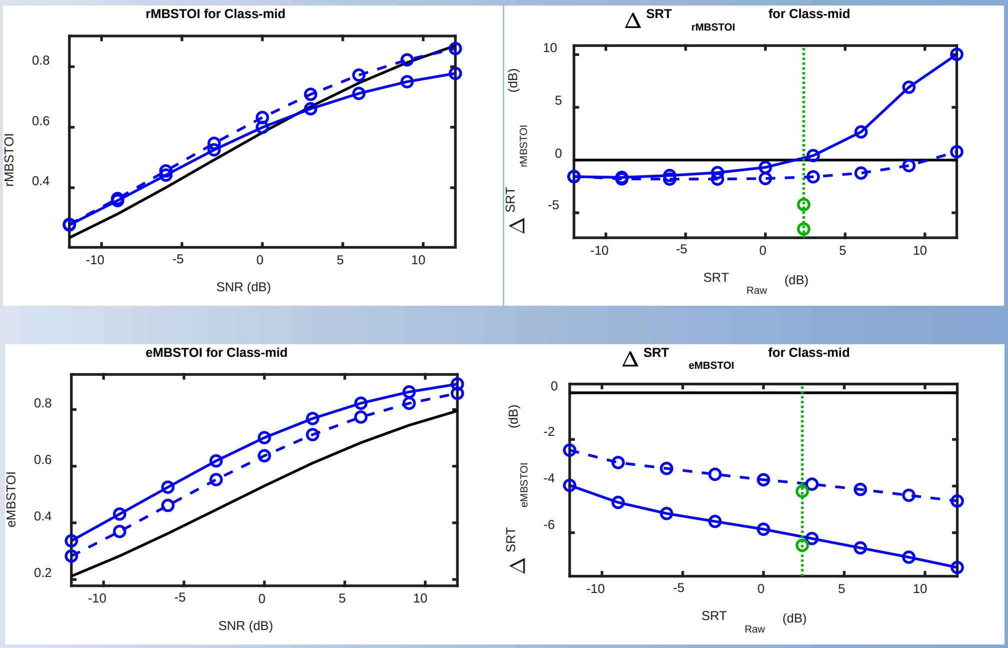
[Moore et al, 2018]

- Classroom full of noisy children. Highly non-stationary.
- Talker = loudspeaker, Listener = KEMAR head/torso simulator.
- MVDR beamformers:
  - Bilateral (2 mic): preserves spatial cues of noise sources
  - Binaural (4 mic): higher SNR, collapses noise to target direction
- Enhancement applies a time-frequency gain:
  - Common gain preserves binaural cues
  - Max function ≈ “better ear”

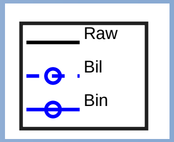
Metric Reference Alternatives

E-Lobes

- MBSTOI needs a clean speech reference:
  - Upper plots use reverberant clean speech as reference.
  - The green o shows the Δ median-SRT @ 50% for 17 HI listeners.
  - Lower plots use the early room response (50 ms) to create the reference.



[MBSTOI: Andersen et al, 2018]

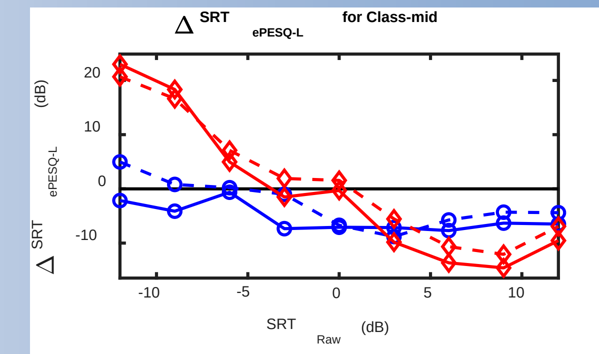


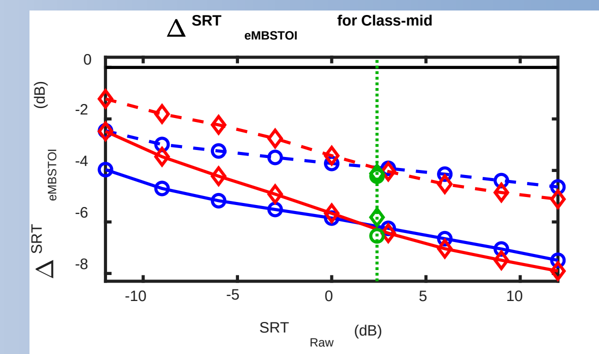
- When reverberant clean speech is used as the reference:
  - MBSTOI predicts small gains that do not match reality
  - Wrongly predicts that bilateral beamformer is better than binaural
- When early part of room response is used to create the reference:
  - MBSTOI correctly predicts ΔSRT for both bilateral and binaural beamformers

Bilateral versus Binaural

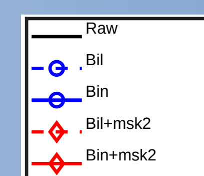
E-Lobes

- Binaural (solid lines) is always better than bilateral (dashed) for both PESQ and MBSTOI
- Enhancement, ♦, improves PESQ and MBSTOI for SRT_{Raw}>2.5 dB but degrades them below this.
  - Worse than the single-channel results





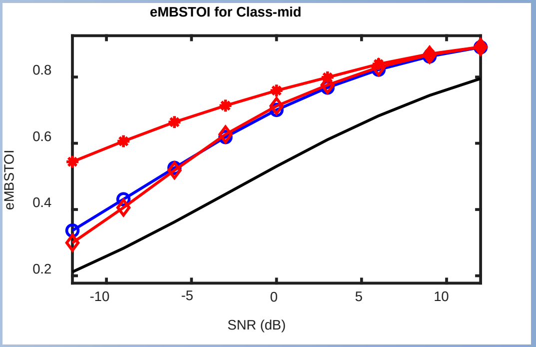
- Measured performance, ●♦, of HI listeners shows that enhancement, ♦, degrades median SRT of binaural beamformer, ●, by 1 dB.

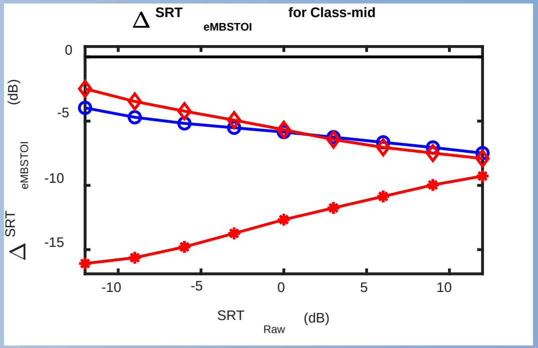


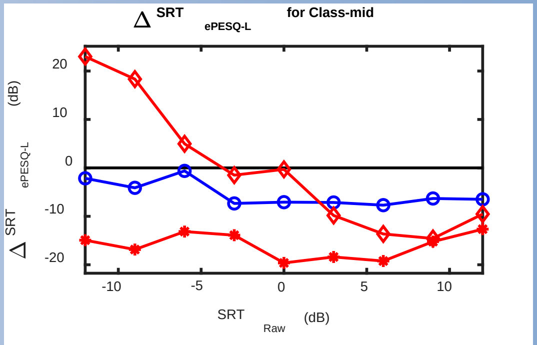
Effect of Better Mask

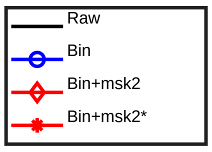
E-Lobes

- Effect of using a better mask (* plot)
  - Fix the mask as the one determined for +12 dB SNR
  - MBSTOI declines more slowly with decreasing SNR
  - ΔSRT_{MBSTOI} continues to improve as SNR decreases
  - PESQ is improved at all SNRs
- Mask-assisted MMSE enhancement can give excellent results with a good enough mask









Summary

E-Lobes

- Mask estimation
  - Aims to identify time-frequency cells that have high speech energy rather than high SNR (maximize STOI of vocoded noise)
  - Depends only on the target source and is single-channel
- Clean-speech reference for metrics
  - Metrics should use a non-reverberant clean-speech reference
  - Useful to express metric in terms of ΔSRT
- Binaural versus Bilateral
  - For noise without dominant point sources, binaural $\gg$ bilateral
  - Better SNR outweighs spatial cue preservation
- Mask-assisted LogMMSE enhancement
  - Can give significant gains but needs a better mask estimator

References

E-Lobes

- Andersen, A. H., de Haan, J. M., Tan, Z.-H. & Jensen, J. (2018), ‘Refinement and validation of the binaural short time objective intelligibility measure for spatially diverse conditions’, *Speech Communication* **102**, 1–13.
- Ephraim, Y. & Malah, D. (1985), ‘Speech enhancement using a minimum mean-square error log-spectral amplitude estimator’, *IEEE Trans. Acoustics, Speech and Signal Processing* **33**(2), 443–445.
- Gonzalez, S. & Brookes, M. (2013), Speech active level estimation in noisy conditions, in ‘Proc. IEEE Intl Conf. Acoustics, Speech and Signal Processing’, Vancouver, pp. 6684–6688.
- Kjems, U., Pedersen, M. S., Boldt, J. B., Lunner, T. & Wang, D. (2010), Speech intelligibility of ideal binary masked mixtures, *in* ‘Proc. European Signal Processing Conf.’, Aalborg, Denmark, pp. 1909–1913.
- Lightburn, L. & Brookes, M. (2015), SOBM - a binary mask for noisy speech that optimises an objective intelligibility metric, *in* ‘Proc. IEEE Intl Conf. Acoustics, Speech and Signal Processing’, Brisbane.
- Lightburn, L. & Brookes, M. (2016), A weighted STOI intelligibility metric based on mutual information, *in* ‘Proc. IEEE Intl Conf. Acoustics, Speech and Signal Processing’, Shanghai.
- Moore, A., Lightburn, L., Xue, W., Naylor, P. & Brookes, M. (2018), Binaural mask-informed speech enhancement for hearing aids with head tracking, *in* ‘Proc. Intl Wkshp Acoustic Signal Enhancement’, Tokyo.