Mask-assisted speech enhancement for binaural hearing aids

ELOBES2019 workshop – 12 January 2019

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

• 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)

- Additive noise

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

- 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
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

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

- 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

- LogMMSE enhancer assumes zero-mean complex Gaussian speech and noise STFT coefficient distributions
  - Gain function depends on posterior SNR, $\gamma$, and prior SNR, $\xi$
- 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, $\xi$
Single-channel Enhancement

- 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 \(\Delta SRT\) of \(-1.5\) dB
- In contrast, LogMMSE enhancer has \(\Delta SRT\) 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

- 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 $\approx$ “better ear”
Metric Reference Alternatives

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

- 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 $\Delta$SRT for both bilateral and binaural beamformers
Bilateral versus Binaural

- Binaural (solid lines) is always better than bilateral (dashed) for both PESQ and MBSTOI
- Enhancement, ♦, improves PESQ and MBSTOI for $SRT_{\text{Raw}} > 2.5$ 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

- 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
  - $\Delta SRT_{MBSTOI}$ 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

• 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 $\Delta SRT$

• 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