BINAURAL SPEECH SOURCE LOCALIZATION USING TEMPLATE MATCHING OF INTERAURAL TIME DIFFERENCE PATTERNS

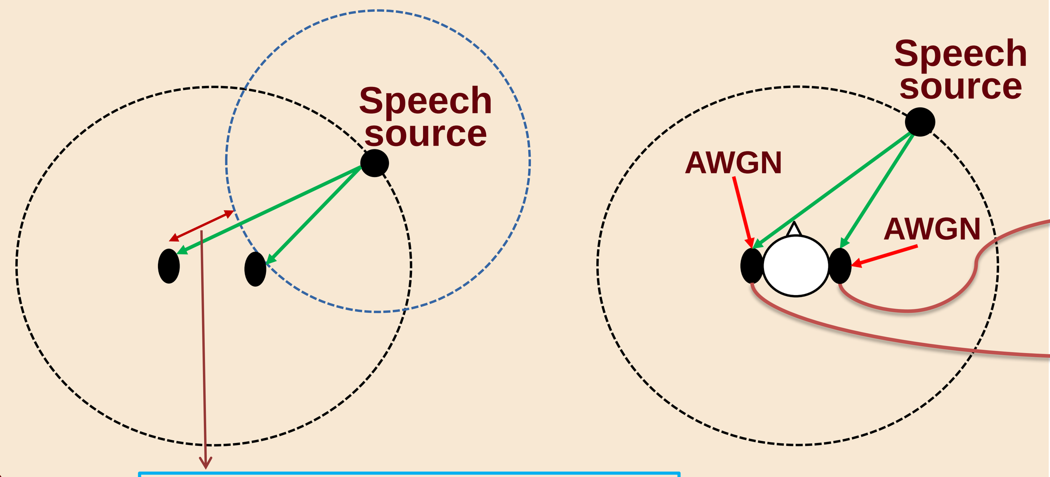
Girija Ramesan Karthik, Prasanta Kumar Ghosh

SPIRE Lab, Electrical Engineering, IISc, Bengaluru.

Objective

- Machine localization of sound sources is necessary for applications such as human-robot interaction, surveillance and hearing aids.
- Adding more microphones can help increase the localization performance. However, humans have an incredible ability to localize sounds with just two ears using two major cues i.e, interaural time difference (ITD) and interaural level difference (ILD).
- Our objective is to localize a speech source from a binaural recording using ITDs and propose a new method using template matching of ITD histogram.

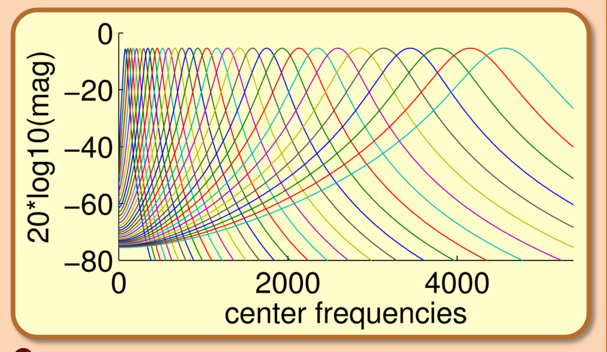
Localization setup



- This difference in the distance travelled causes interaural time difference (ITD) and level difference (ILD) between the two microphone signals.
- Given omnidirectional microphones, an impulsive source will have frequency independent ITD & ILD.
- However, in binaural recordings, reflections & diffractions caused by the head makes ITD & ILD frequency dependent. This dependency is captured by the Head Related Transfer Function (HRTF).

Frequency dependent ITD extraction

Gammatone filterbank



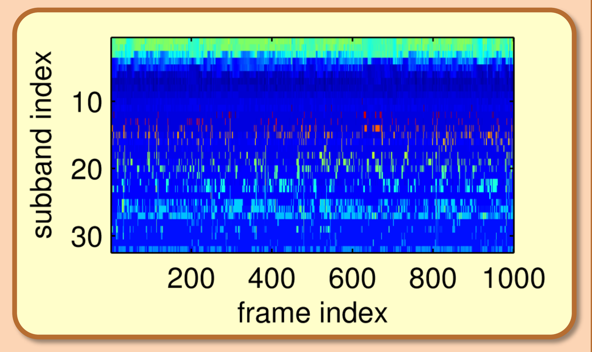
- Approximates the human auditory filterbank.
- 32 equally spaced subbands w.r.t. the Equivalent Rectangular Bandwidth (ERB) scale are considered between 80Hz and 5kHz.
- This range covers most of the speech energy.

Left ear

32 filtered subband outputs

Right ear

ITD spectrogram



- ITD is estimated as the delay with the maximum cross-correlation between the left and right subband outputs.
- ITD (τ) is estimated in each of the nf frames & ns subbands to obtain the spectrogram shown above.

Localization

Maximum Likelihood (ML) based localization using trained GMMs

Obtain likelihood of ITDs w.r.t. each direction (k) in each frame (j) in each subband (i)

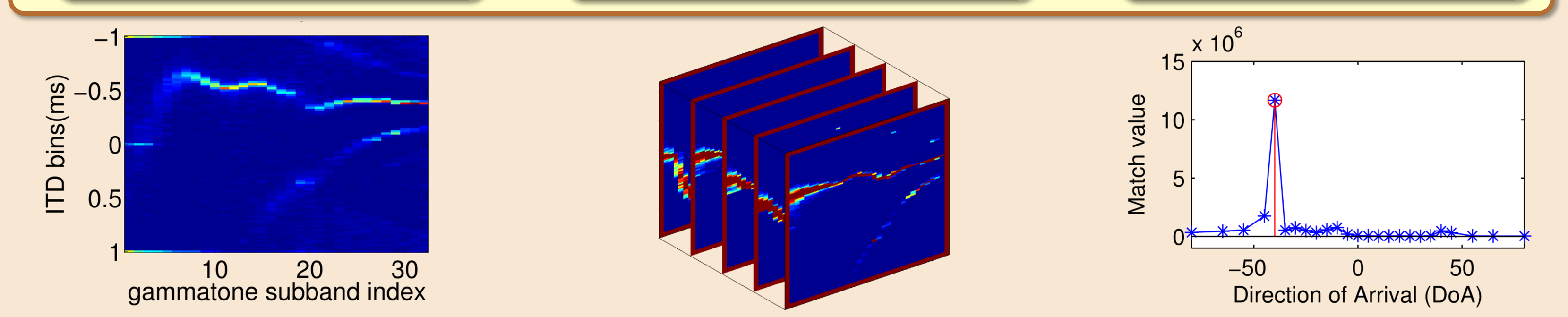
Binaural Hist [1]: $\hat{\theta}(j) = \underset{\theta_k}{\text{argmax}} \sum_{i=1}^{ns} \log(P(\tau_{i,j}|\theta_k, i))$ → Mode of the frame level DoA estimates $\hat{\theta}(j)$ → DoA

Binaural ML [2]: $\hat{\theta} = \underset{\theta_k}{\text{argmax}} \sum_{j=1}^{nf} \sum_{i=1}^{ns} \log(P(\tau_{i,j}|\theta_k, i))$ → DoA

GMMs are trained on the ITDs of each direction for each subband.

Proposed template based localization

Generate test template (IPT) → Match test IPT with the Reference IPTs → Pick the peak from the match profile → DoA



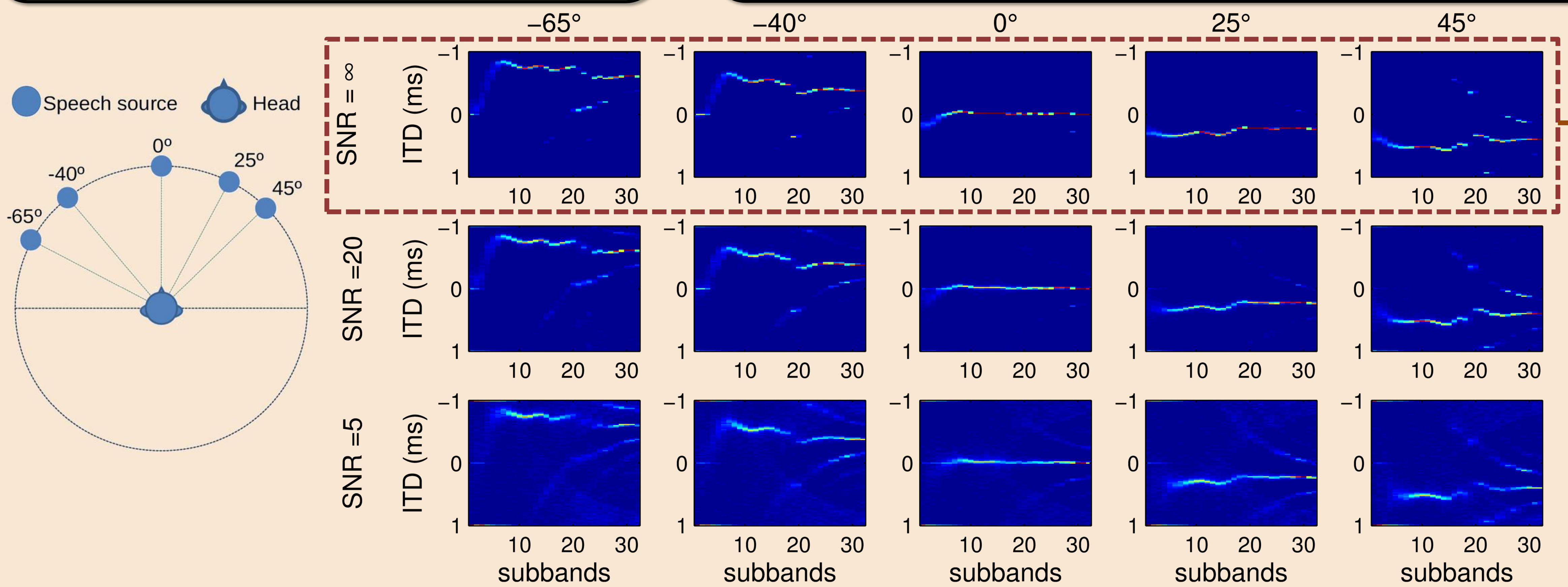
ITD Pattern Templates (IPTs)

Generation: IPT is obtained by stacking the ITD histograms from each subband. The figure below shows IPTs for different directions at various SNRs. All the IPTs shown here are obtained using nf = 1000 i.e., a duration of 10s.

$$k^* = \underset{k}{\text{argmax}} \sum_{i=1}^{ns} \sum_{b=1}^{nb} (T_{test}(b,i) \times T_k(b,i)) \Rightarrow \hat{\theta} = \theta_{k^*}$$

Matching: Match between test IPT T_{test} and T_k (reference IPT for the k^{th} direction) is given by sum of the elements of their Hadamard product.

Localization: Picking the direction with the maximum match.



Clean IPTs are used as the reference IPTs as they encode the direction dependent patterns without the interference of any noise or reverberation. These are obtained by convolving speech with the HRTFs from the CIPIC database [3].

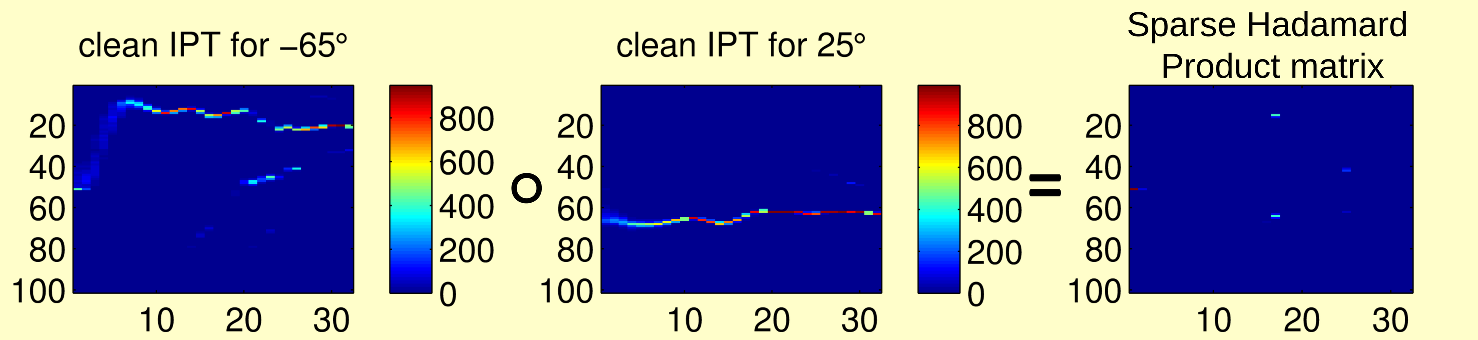
Complexity of IPT generation

IPT generation involves histogram computation. Given the lower limit (l) and bin width (bw), the bin index $i(v)$ associated with a data point v is given by the equation below

$$i(v) = \left\lceil \frac{v - l}{bw} \right\rceil$$

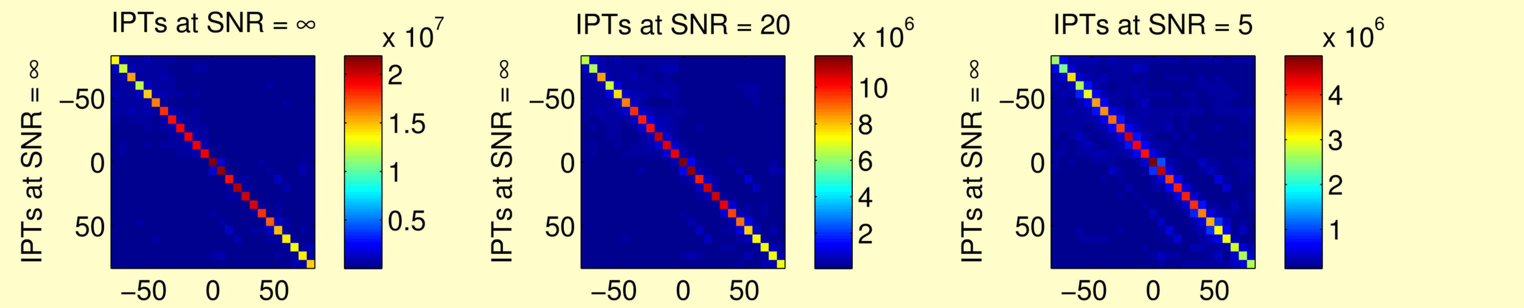
Hence, the histogram computation is $O(n)$.

Reference templates are almost non-overlapping



It can be seen above that the Hadamard product (○) of reference templates of two different directions form a sparse matrix, suggesting that the patterns are almost non-overlapping.

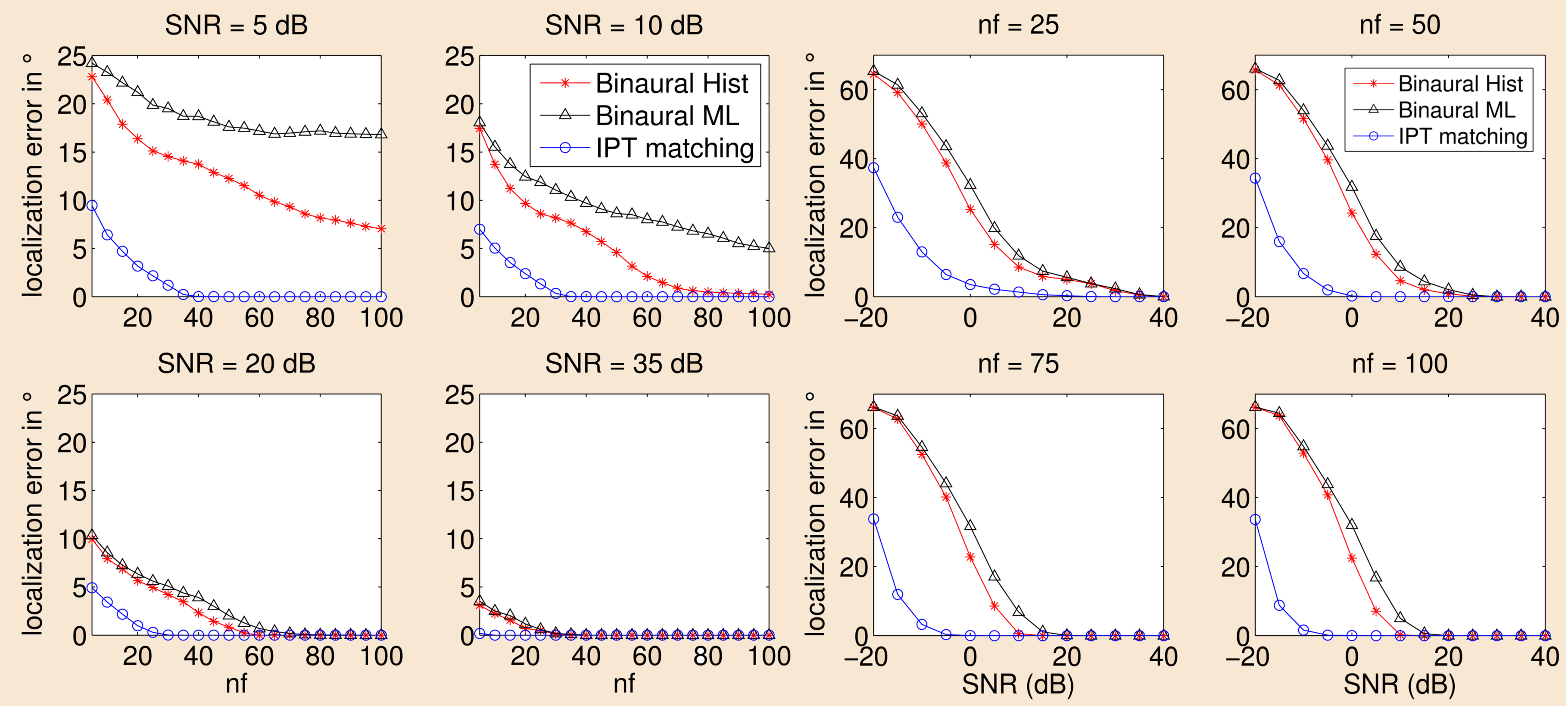
Similarity/Match matrices between reference templates (SNR = ∞) and templates at different SNRs with AWGN



The similarity is calculated using the match measure defined above. This shows that, under additive white Gaussian noise (AWGN), the peak still occurs on the diagonal signifying correct localization in the presence of AWGN.

Results

nf – Number of frames used for localization



Conclusion

- A new template based localization algorithm has been proposed using templates (IPTs) generated from ITDs under anechoic conditions.
- The patterns in clean IPTs are well preserved under additive white Gaussian noise (AWGN). This validates the use of clean IPTs for localization under AWGN conditions.
- An $O(n)$ method to compute IPTs makes it computationally efficient.
- As part of further analysis, we would like to extend the use of IPTs to reverberant and multiple speech source scenarios.

References

[1] T. May, S. van de Par, and A. Kohlrausch, "A probabilistic model for robust localization based on a binaural auditory front-end," IEEE Transactions on Audio, Speech, and Language processing, vol. 19, no. 1, pp. 1–13, 2011.

[2] John Woodruff and DeLiang Wang, "Binaural localization of multiple sources in reverberant and noisy environments," IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 5, pp. 1503–1512, 2012.

[3] V. R. Algazi, R. O. Duda, D. M. Thompson and C. Avendano, "The CIPIC HRTF database," in IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics, pp. 99-102, 2001.

Authors thank the Pratiksha Trust for their support

{karthikgr, prasantg}@iisc.ac.in