We would like to thank the reviewers for their thoughtful comments.

General comments:
Train/test data split, generalization to real results (R3, R4): Our test and train sets are completely independent and there are no overlapping identities between train/test splits—our network is *only* trained on VCTK data but performs well on real speakers in unseen environments, as demonstrated in the supplementary video. As requested, we ran numerical results on a test set of real recorded data, and found a median SI-SDRi of $10.7\,\mathrm{dB}$ on mixtures of 3 speakers, higher than Oracle IBM $10.2\,\mathrm{dB}$ on the same data. Full comparisons and results will be reported in the final version.
Elevation Angle (R1, R4): Many real-world situations, like the ones in the supplementary video, can be well approximated by an azimuthal-only model. Although our work only focuses on localization and separation by azimuthal angle, we will add the note by R1 about our model's assumption regarding the relationship between elevation angle and TDOA. We will additionally add a discussion on possible extension of the proposed model to handle elevation.
Experimental data (R1, R3): Full details on rendering parameters will be added in the final-version.

Reviewer #1 (R1):
Eq. 6: $\mathbf{s}_i'$ is computed by using Eq. 3 the same way that $\mathbf{x}'$ is computed, using the TDOA.
Conv-TasNet: We modified the single-channel Conv-TasNet to handle multi channels by changing the number of input channels and output channels on the first and last layer of the network.
Is the method expected to generalize to any array geometry?: We need to use a different model trained on each different array geometry. **Post-processing**: All synthetic results and numbers are reported without post-processing. We will add qualitative real results to show the comparison.

Reviewer #2 (R2):
Beamforming: We thank the reviewer for pointing out the analogy to tunable-width beamforming, which we will add in the related work. The benefit of our method over tunable-width beamformers is an ability to select a certain type of audio (e.g., speech), which is otherwise impossible to separate if the speaker is spatially close to a noise source. We will also add a comparison to state-of-the art beamforming in the experiments section as suggested.
Global Sweep and Tracking: Although we use binary search as a global sweep, our method can also be used as suggested to perform local sweeps in subsequent time steps. We will add a discussion of this.
Usefulness and Speed: It is true that the network is not specifically designed for real time processing, but it can run on chunks as small as 0.5 s, making it practical for many use cases such as smart home devices, media production, or meeting transcriptions. Real time processing is an exciting future work, which could use slightly different network architectures while maintaining our core idea of a target angle and variable window size.

Reviewer #3 (R3): Please confirm this review is for our paper since the summary is for a lip sync paper.
Different sampling rate: When running at $16\,\mathrm{kHz}$, we found that the median angular error was worse by $1.3°$, while the separation results were worse by $2.23\,\mathrm{dB}$. We can add these lower sample rate experiments and comparisons to SOTA in the paper. We believe that the ability to operate at higher sample rates is a major benefit of our method that will allow extensions to other source types, for instance, high-resolution music signals.
Clean target audio: When we say "clean sample," we just mean we need each source separately since our method is supervised. **Why cannot the samples be mixed with reverberations?** The target audio is mixed with reverberations, which is then used as the ground truth target. More clarifications will be included in the final-version.
Oracle IRM: Yes, the low result is correct. We used the evaluation code provided as a part of SiSEC 2018 campaign, and confirmed qualitatively that IRM is poor. Because the mixtures contain high levels of background noise, the IRM contains mostly phase information from the background. **Is post-processing used in synthetic testing?** No, it is not.

Reviewer #4 (R4):
Unrealistic simple signal model and experiments: We would like to correct some misunderstandings regarding our rendering method. Signal attenuation and reverb are both considered and modeled in the synthetic data with the `pyroomacoustics` library. Furthermore, both real and synthetic results show that phantom sources are not picked up.
Leakage of Source: The variable window size is motivated by the non-point source nature of sources. Although the synthetic data is modeled as point sources, the real results show that we can handle moving and non-point sources. We also show localization and separation on 4 real sources, not just 2.
Localization and separation results only on simulated data: We note that many recent papers on spatial audio such as [1] from 2020 and [2] from 2018 show results only on synthetically rendered data.
"Not convincing that this approach can be applied in practice": The real results we have shown are not cherry picked, and are natural scenarios with speakers and environments not seen during training. This represents evidence that it can indeed work in practice. In addition, our real results are actually shown on up to 4 real speakers, not only 2.

[1] Luo, Yi, et al. "End-to-end microphone permutation and number invariant multi-channel speech separation." *ICASSP*. 2020.
[2] Johnson, Daniel, et al. "Latent Gaussian activity propagation: using smoothness and structure to separate and localize sounds in large noisy environments." *NeurIPS*. 2018.