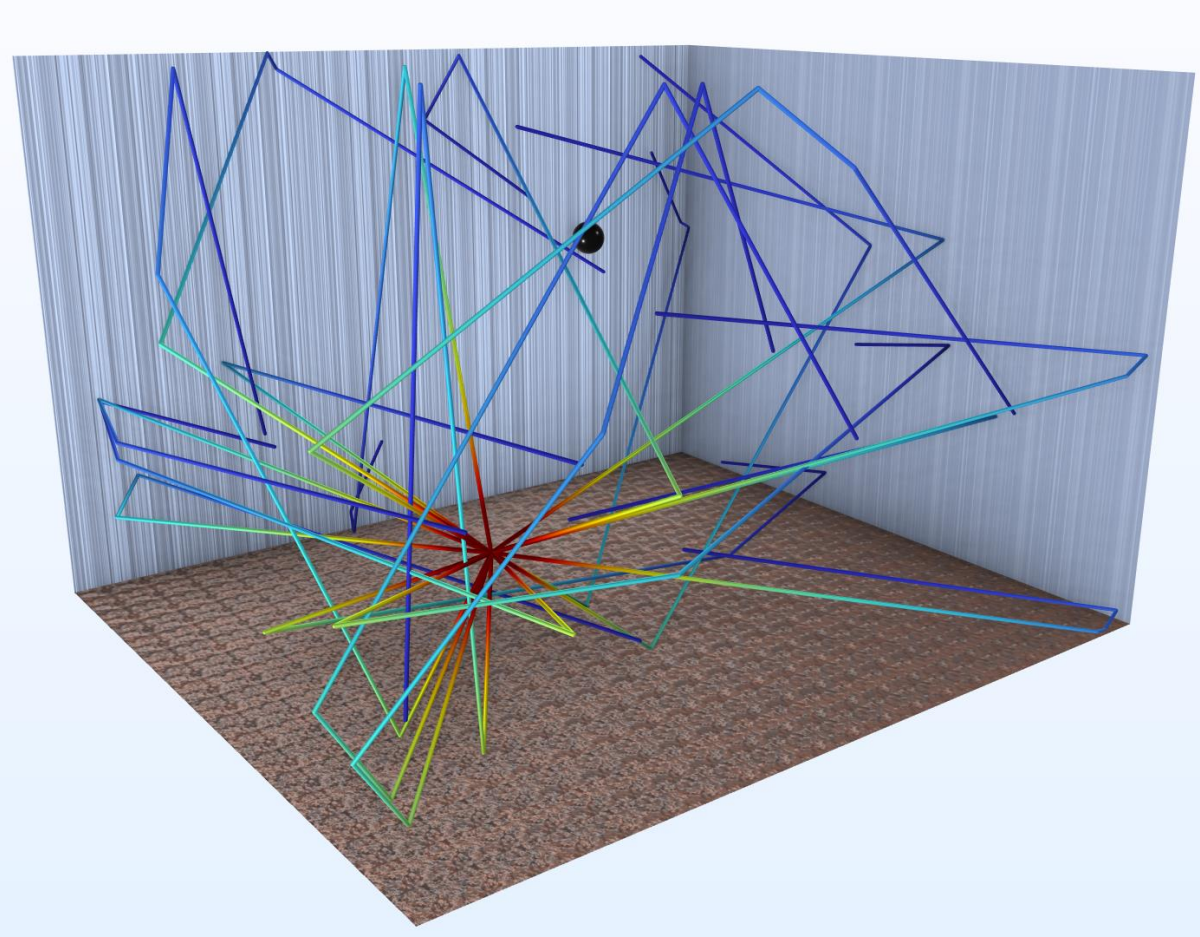
INDUCTIVE BIASES FOR NOVEL VIEW ACOUSTIC SYNTHESIS (NVAS)

A LOOK INTO CLASSICAL METHODS

HOW RIRS WORK

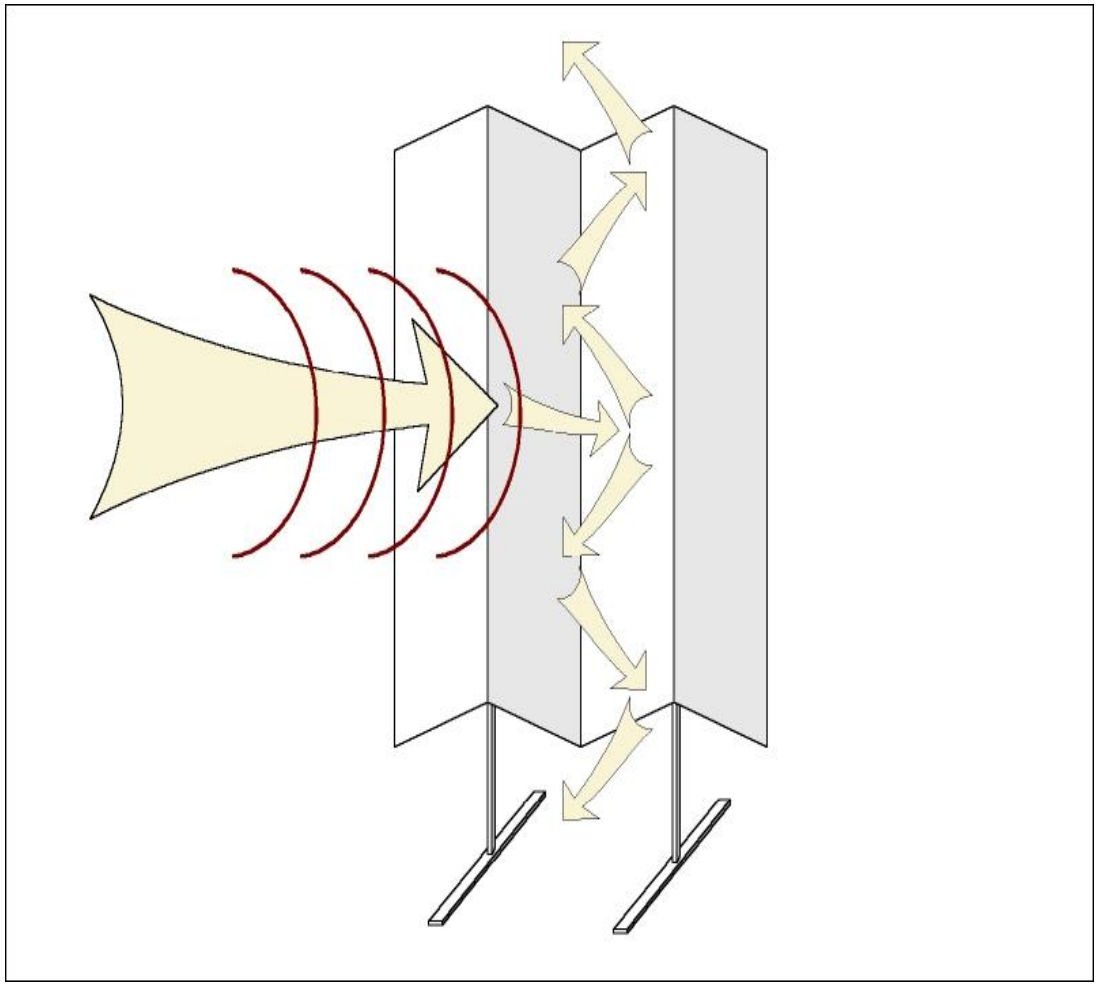
EARLY REFLECTIONS (0-50/80MS)

- Mostly modeled via the classic methods of Ray Tracing and Image Source Method (ISM)
- The early reflections are mostly formed by arrival time of each ray, so the more "bounces" a ray has to do to arrive at destination, the later it is reflected in RIR
- Mostly geometric and frequency independent
- Much more sensitive to time of arrival for each ray
- Phase of the sound is important and calculable in this part
- What matters most in this region is that which rays arrives when



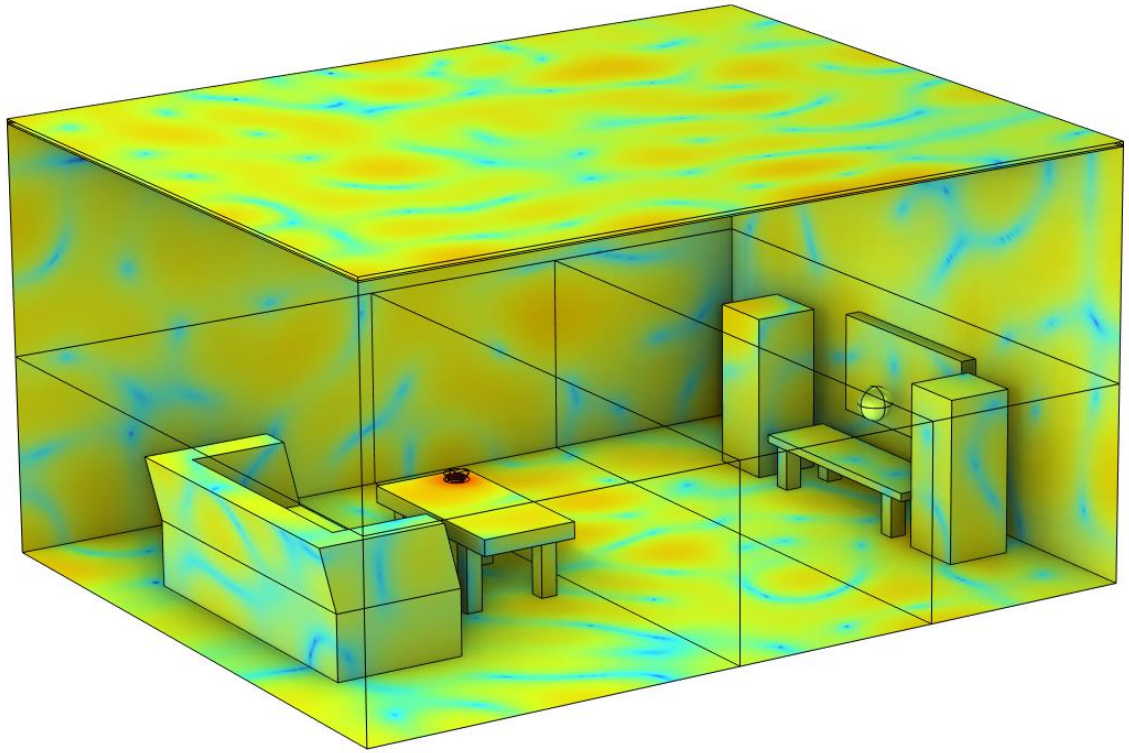
MIDDLE PART / TRANSITION OF RIR (50/80MS – 120/150MS)

- As rays bounce around, they scatter energy and get more "diffuse"
- Gradually they lose their geometric property and become more random
- Higher frequencies lose their energy more while lower frequencies hold the energy better
- This region is geometric for rays that scatter less or bounce of reflective surfaces, while becomes more diffuse if rays bounce more or bounce off rough surfaces
- No clear classic model for this part



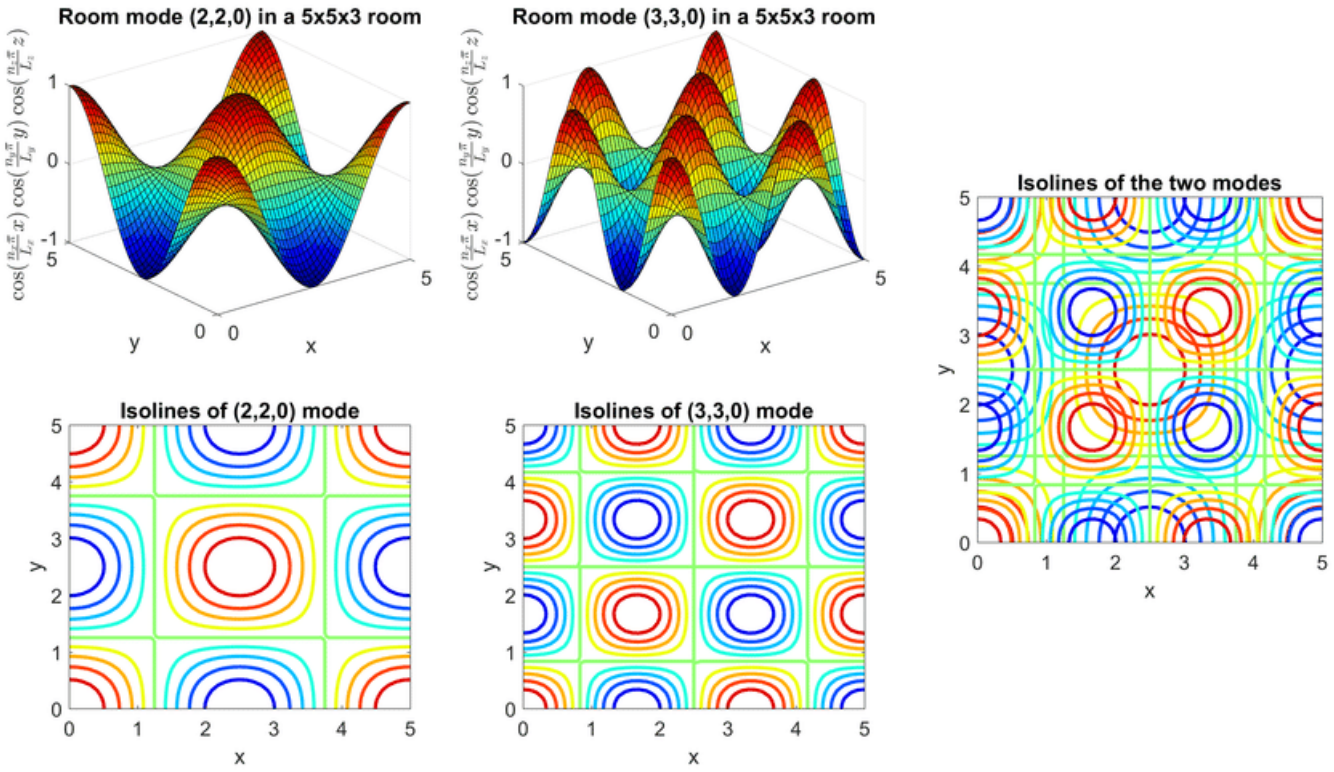
LATER REVERBERATIONS (120/150MS – END OF RIR)

- After some time all rays are diffuse, are much more random and phases get fully random and meaningless
- Energy is mostly like a field in 3D space of room that is exponentially decaying
- This field of energy and how it looks is mostly dependent on how early reflections were and consequently, room shape and source position, After around 300ms it is independent of even source position
- One of the classic models to represent this part is Polack model, where it decays an initial random noise using per frequency band exponentials. Another more accurate model is FDN which mixes multiple samples from previous signals to calculate next sample.
- In this part, high frequencies usually lose their energy while lower frequencies endure.
- Unlike Early Reflections, late reverberations can be modelled decently with autoregressive models



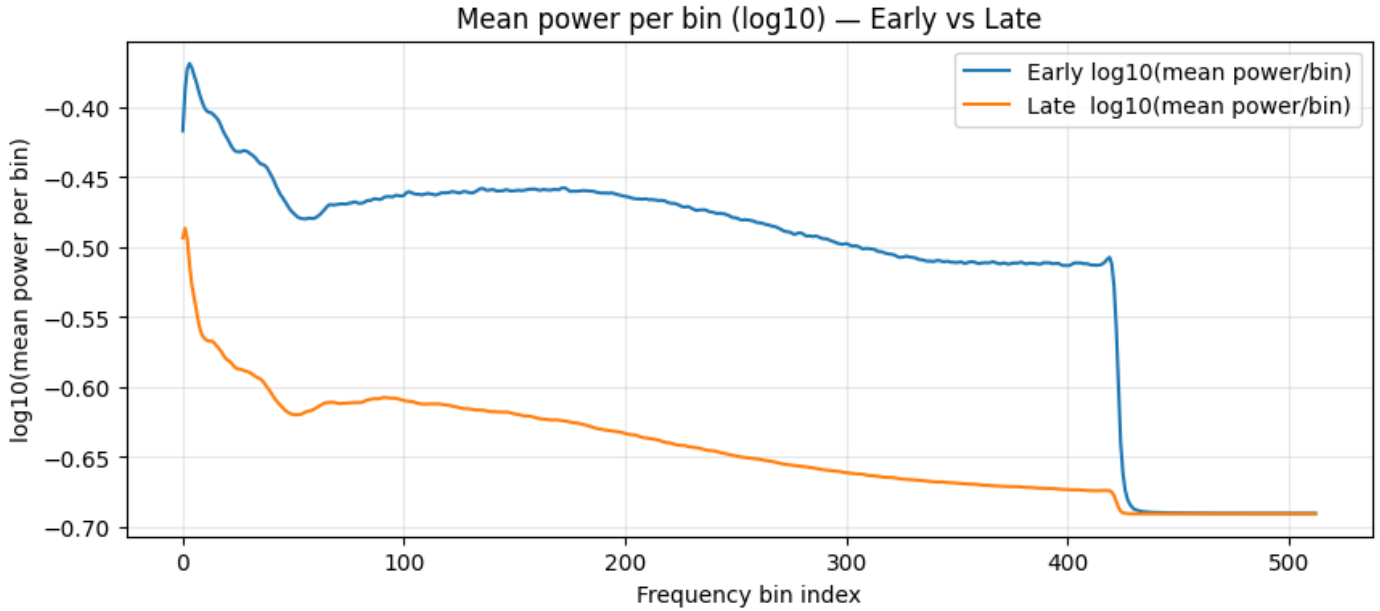
ROOM MODES

- Below a certain frequency (usually 300hz), sound does not behave like rays bouncing around and The room starts acting like a resonant box
- Certain frequencies "fit" the room shape and ring more than others which are called the room modes
- Some modes are less damped, so once excited they persist longer.
- Below certain frequencies, it's mostly about which modes persist rather than exponentially decaying noise.
- Because low-frequency modes are typically less damped, they dominate the late part of the energy decay curve and contribute disproportionately to late reverberation energy.
- Modes are modal and not geometric for both early reflections and late reverberations



ROOM MODES

- The image shows amount of energy in both ER and LR per frequency band
- Low frequencies (room modes), have highest amount of energy in both parts
- In LR, the amount of energy they contribute is more substantial



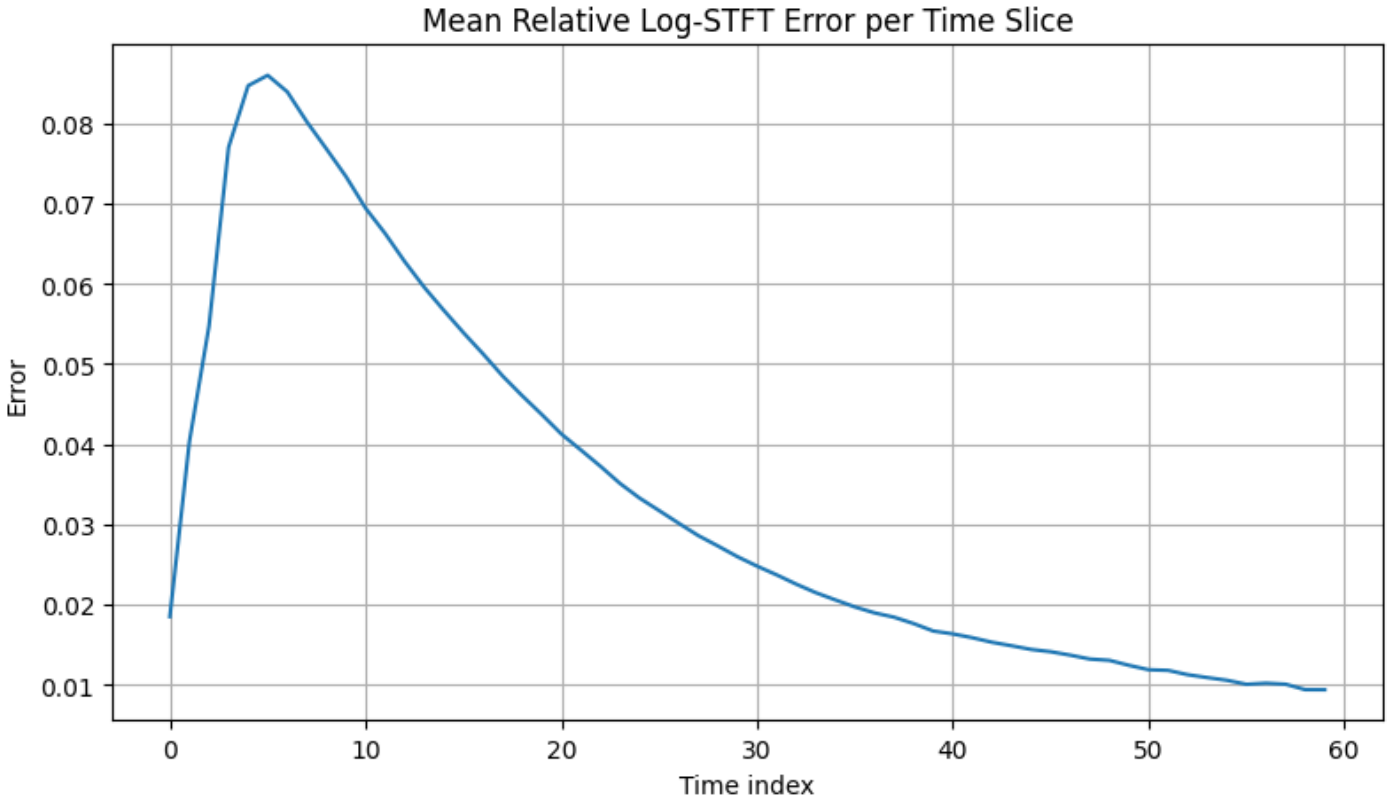
KEY TAKEAWAYS

- Early reflections are time sensitive and geometrics
- Late reverberations depend heavily on past signals, and are more statistic in nature
- Low frequencies are mostly behaving differently from every other frequency band in both ER and LR
- These 3 parts are naturally different, in mathematics and modelling

CURRENT NVAS SYSTEM FAILURE ANALYSIS

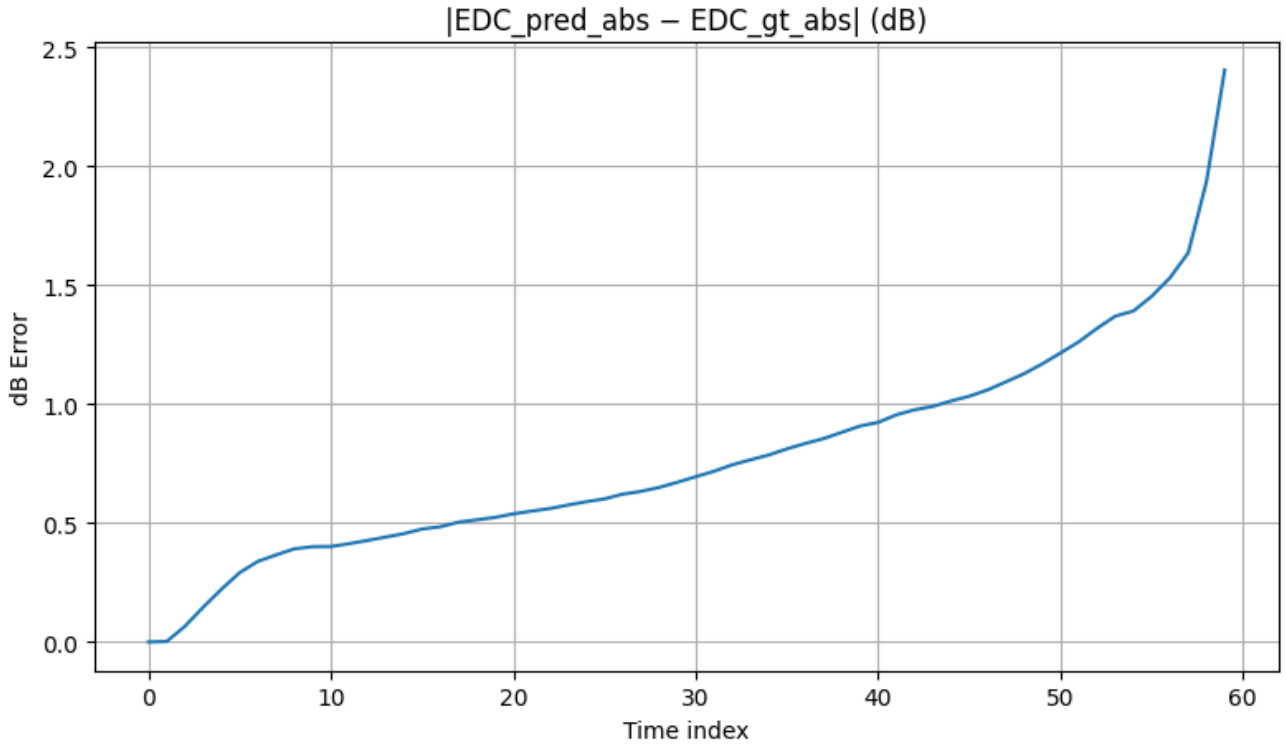
TEMPORAL ERROR (EARLY REFLECTIONS)

- Each time bin is 5ms
- Most of the error is in first few time bins, meaning it is in early reflections part (ER)
- The system can't fully learn geometric properties of the room
- In ER, it is highly time dependent, while in LR it is mostly general decay over time, not exact sensitivity to time
- ER is phase dependent, unlike LR



DECAY ERROR (LATE REVERBERATIONS)

- Steadily increases over time
- Since it depends on past frames, predicting frame by frame reduces reliability of curve prediction
- This part is neither sensitive to frequencies nor time, but mostly sensitive to past data
- Since model has modal understanding issues, a descent portion of EDC error comes from modal part
- Decay rate is different for each frequency, higher frequencies decay faster than lower frequencies



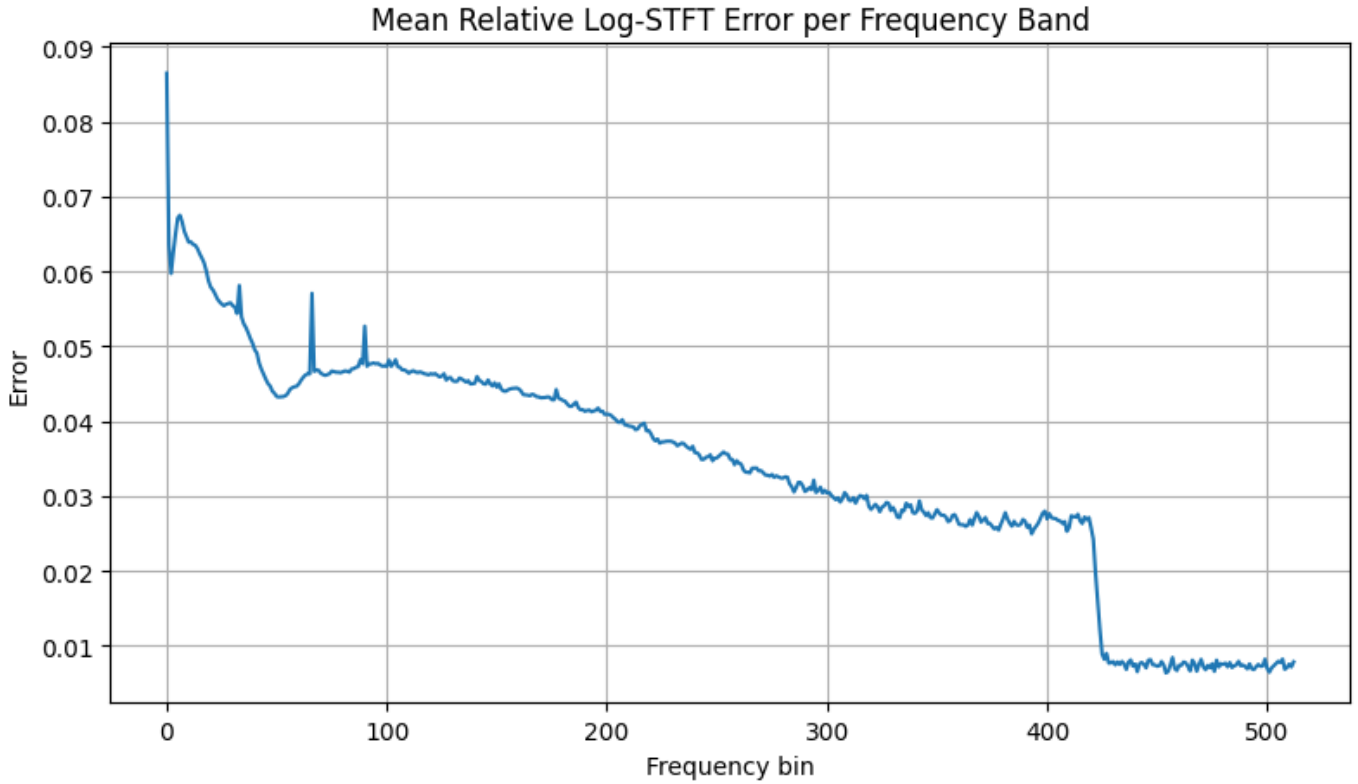
FREQUENCY ERROR (MODAL STRUCTURE)

- Each frequency bin is 38hz
- Most of the error is in low frequencies
- The system can't fully learn modal properties of the room, which is defined as:

$$h(t) \approx \sum_{k=1}^{K} A_k \, , e^{-\alpha_k t} \cos(\omega_k t + \phi_k)$$

Which is basically sum of dampened sinusoids

- This part is highly sensitive to exact frequencies
- In low frequencies it is mostly about global geometry.
- Tests shown in low frequencies, the system is highly low rank.
- Phase matters but less than ER



LOW RANK STRUCTURE OF ROOMS AT LOW FREQUENCIES

- The STFT was cropped for higher frequencies (basically a low pass filter)
- All time bins across all samples in dataset were concatenated to a single matrix
- The SVD from that matrix was calculated.
- Unlike full STFT which was high rank, these matrices were low rank.
- For full 513 bin STFT:
  - `Rank for 90% variance: 5 Rank for 95% variance: 57 Rank for 99% variance: 167`
- For full 50 bin STFT:
  - `Rank for 90% variance: 4 Rank for 95% variance: 11 Rank for 99% variance: 23`
- For full 20 bin STFT:
  - `Rank for 90% variance: 3 Rank for 95% variance: 5 Rank for 99% variance: 10`

INDUCTIVE BIAS IDEAS FOR EARLY REFLECTIONS

1) Having an auto-regressive model
2) Having more supervision to handle time sensitivity in first 50-80ms, main ideas being either of the auxiliary losses:
    1) A multi window size STFT loss that has different temporal resolutions
    2) A loss on waveform matching
3) Predicting phase alongside magnitude as it directly impacts time of arrival (Since phase is discontinuous, it's tough to predict, NAF paper uses instantaneous frequency (IF) to predict phase)
4) Inputting time not as discrete time query tokens, but continuous values, with continuous positional encodings

INDUCTIVE BIAS IDEAS FOR LATE REVERBERATIONS

1) Having an auto-regressive model
2) Having a loss that handles the "different decay rates per frequency band" better
    1) EDC loss per frequency band
3) Helping to learn room modes better (next slide)

INDUCTIVE BIAS IDEAS FOR ROOM MODES

1) Having a loss to learn better frequency sensitivities:
    1) Learning a multi frequency resolution loss with higher frequency resolution at lower frequencies, but lower temporal resolution
    2) Having an auxiliary DFT loss with low weight.
2) Having another head to predict all of lower frequencies all at once, from a shared token with baseline and this head. Maybe using Low-Rank global basis vectors for this
3) Using physical informed model design to parametrically learn sum of modes as representation for this part:

$$h(t) \approx \sum_{k=1}^{K} A_k \, , e^{-\alpha_k t} \cos(\omega_k t + \phi_k)$$

MAIN IDEAS FOR IMPROVEMENT

1) Have an autoregressive model, best option being self-attention
2) Treat these parts as different, most importantly early reflections (ER) and late reverberations (LR)

MAIN IDEAS

MAIN STRUCTURE

1) Either all learned via a single backbone that is AR
2) Or learn each part via a different head with different loss and properties
3) Have losses for more attention to fine time or frequenency detail

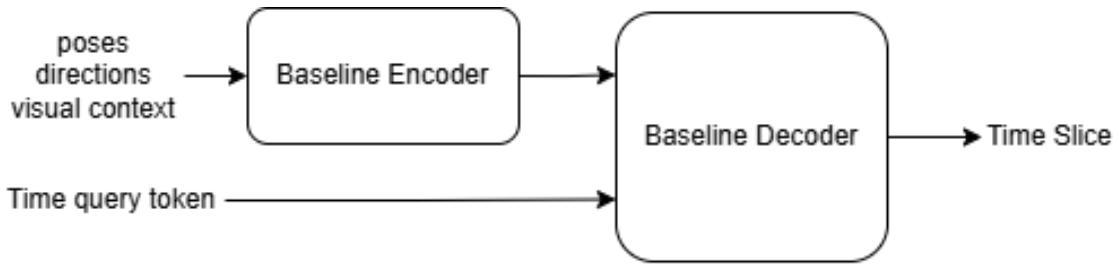
R. Yamamoto, E. Song, and J.-M. Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," Feb. 06, 2020,

AUTOREGRESSIVE ATTENTION IN LATENT SPACE

- Encoder gives 60 tokens for 60 time slices
- A causal attention applies over them, then passed to decoder
- The aim is temporal cohesion

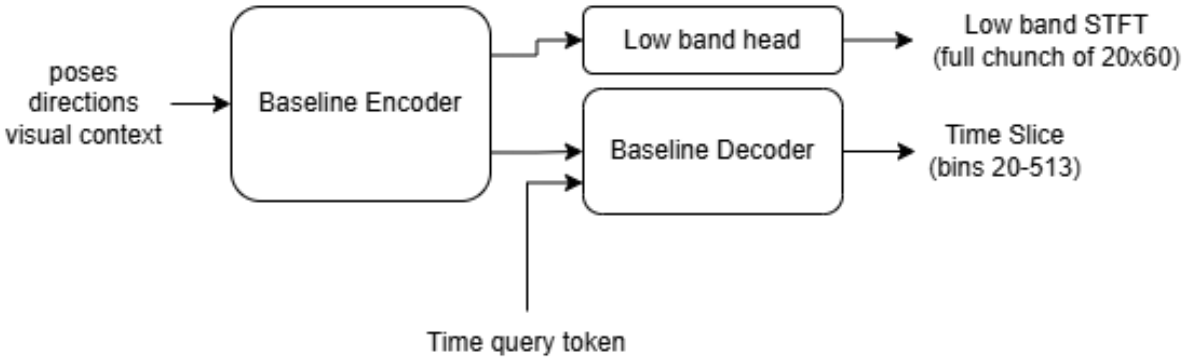
SHARED GLOBAL TOKEN

- This one is mostly a modification that can be used as baseline for multi-head ideas
- Instead of inputting time query token to encoder, we input it to decoder



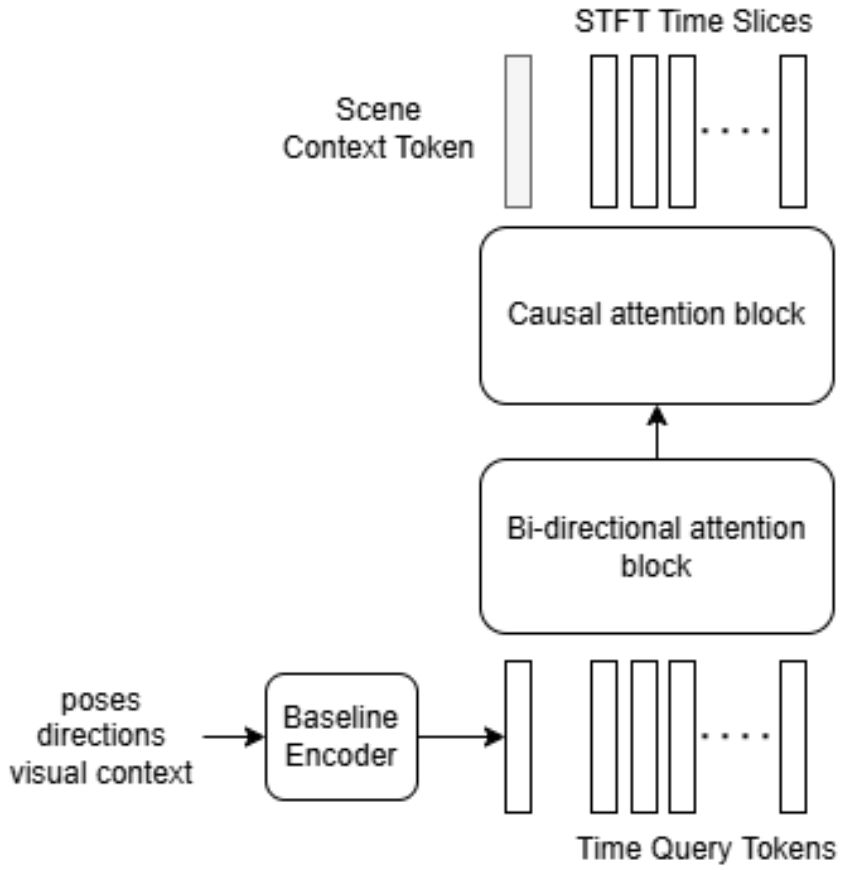
SEPARATE HEAD FOR LOW FREQUENCIES

- Since one of the main fail states and main bottlenecks of model is modal understanding, and modal behavior of room is different, we can separate that part of STFT
- The idea is one head does as normal, one head predicts all of the STFT for modal part (low frequencies) at the same time, while rest of the model predicts one time slice at a time like baseline



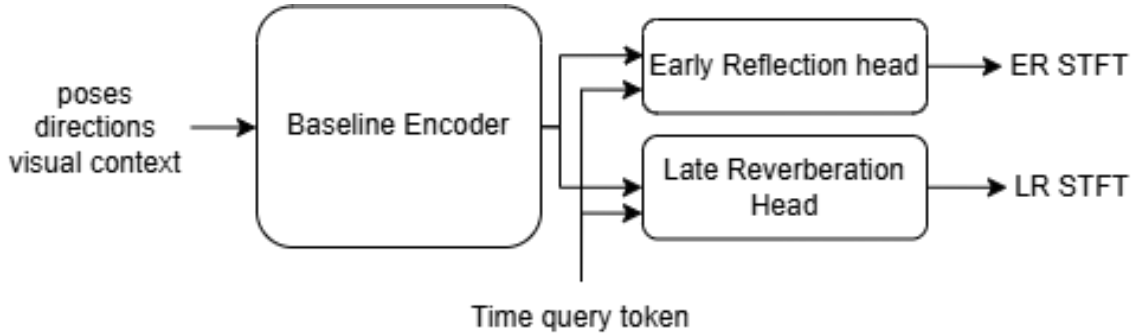
FULL TRANSFORMER ATTENTION

- Instead of applying attention over tokens from baseline, we can just treat all of the RIR generation like a transformer
- Time query tokens + a scene context token are inputted to transformer, then output of it should be full RIR
- This can also be done like "next time slice prediction" but it might be an overkill



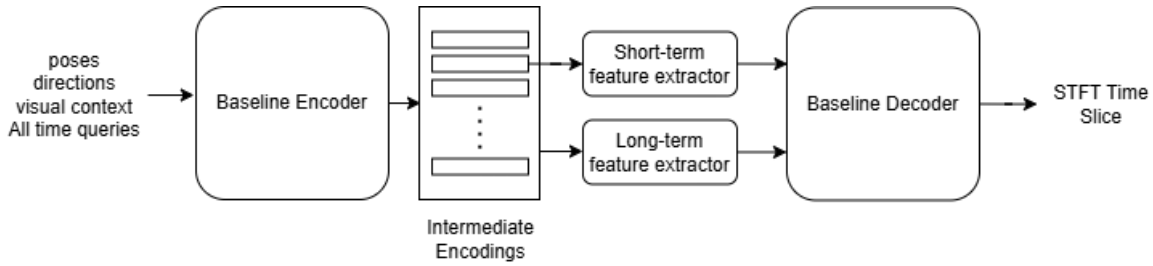
MULTIPLE HEADS FOR ER AND LR

- The main intuition from classic methods is that physically ER and LR are different, so each one is handled by different head with different design
- The head for LR can be more lightweight and more constrained
- Each head might have it's own loss weights, even their own losses
- A problem is the transition between ER and LR, so this idea might not be best to test as first



MULTI TIME SCALE LATENT SPACE

- Instead of applying attention over latent space, we might just extract multiple scale windowed context of different sizes around each center token.
- For example alongside token for time slice t, we pass it feature from a 10 time slice size and full global context into the decoder as well.
- Attention might outperform it, we test this is attention failed



MAIN LOSSES

MULTI BAND EDC

- Intuition from classical methods is that different frequencies decay at different rates
- Instead of global EDC, for each frequency band we apply a different EDC
- This idea was proposed in the paper: Neural Acoustic Context Field: Rendering Realistic Room Impulse Response With Neural Fields

A SMALL MSE WAV LOSS

- While we predict STFT, early reflections are directly related to sensitivities in time and arrival time
- Best representation of audio that captures these fine detail is wav, but on it's own wav is hard to predict and very noisy
- We can apply ISTFT to get wav and with a small weight, calculate MSE loss on WAV.
- This can be done only for ER, also this can be combined with multiple head approaches, with the ER head getting this loss

MULTI SCALE STFT LOSS

- The idea is that we need fine detail temporal for ER and fine detail frequencies for modal structure, but predicting both wav and DFT might be too noisy for the model, so we predict STFT at different scales
- The way it's done is to calculate one STFT, then apply ISTFT by using phases from ground truth to get the wav, then calculate different STFT scales for loss
- A long window STFT with lower temporal resolution but higher frequency resolution is useful for modal structure (it can be weighted more for lower frequencies)
- A shorter window STFT with higher temporal resolution can be applied to help learn ER better, also with higher weight for first 50-80s
- These can be combined with multiple head designs

PREDICTING WAV BUT STFT LOSSES

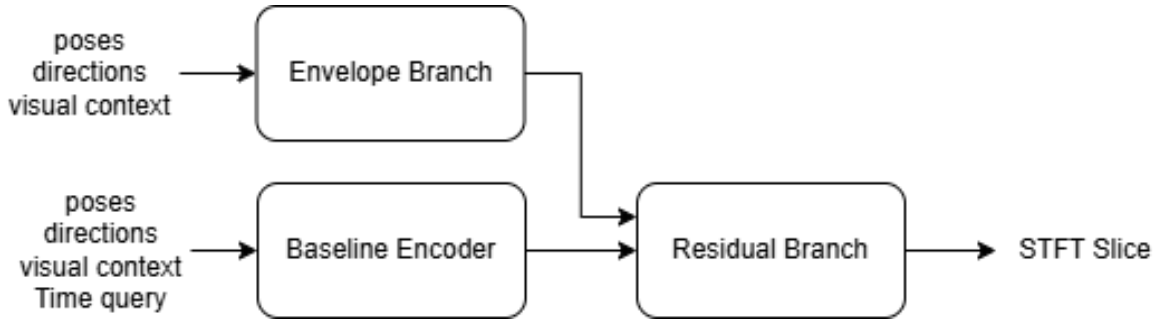
- Same as previous idea but making the model actually predict full WAV, but since wav is noisy, we don't calculate loss in wav, but calculate it in STFT domain
- From the output wav, we predict multiple scales of STFT and calculate loss on these
- This idea is proposed by the heavily cited (1205 citations) paper: Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram
- This idea is tested in speech understanding task and not RIR generation

OBSOLETE IDEAS

THESE IDEAS WERE DROPPED DUE TO FAILING IN SOME ASPECTS

ENVELOPE RESIDUAL DUAL BRANCH

- We tested with predicting global energy, over time then starting with that as baseline for STFT branch and asking it to predict frequency detail, which will then be added to base energy per time slice
- It didn't work because of 2 reasons:
  - Global energy is very rough, as energy is different per band
  - It has non-identifiability issue, with residual branch being very powerful that it overwrites everything we envelope branch predicts

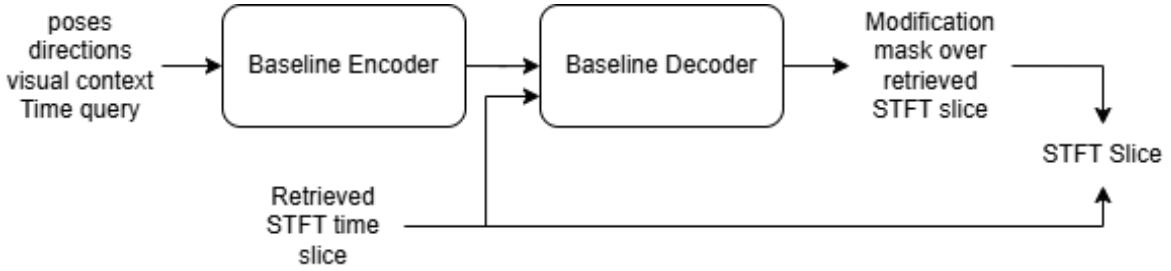


FEATURE LEARNING VIA RAG

- The base RAG might still work but it has some issues that are more fundamental:
  1. The model is not able to have a good temporal modelling, so with retrieved features being better at decay metrics, we can't easily make the model to use them
  2. Retrieved samples are better at decay metrics but worse at STFT error, the problem is that these 2 part are not separable, and if it has bad STFT detail, it would inject them into the model even if we filter for decay metrics
- It might still work if we scaffold some parameters and use them as constraints for RIR generation process, but in my opinion focusing on the fundamental lack of inductive bias might have higher chance of success

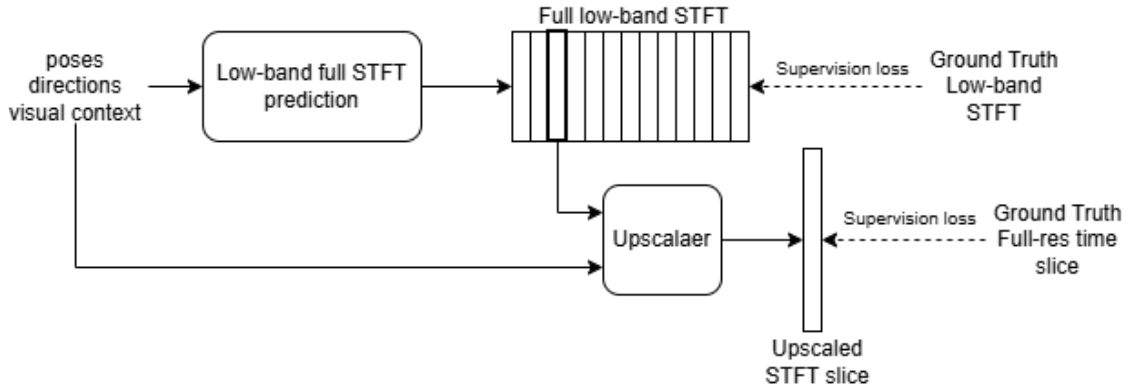
RAG STFT MODIFICATION

- One RAG idea we had but didn't test.
- In this idea, we can retrieve top-1 sample from dataset and use it as base STFT and predict difference mask, modify it to make the STFT error but keep decay metrics the same
- Might still work, but intuition says since decay metrics and STFT error are not inseparable, we might compromise one for the other
- Regardless, this idea might be useful for few-shot setting



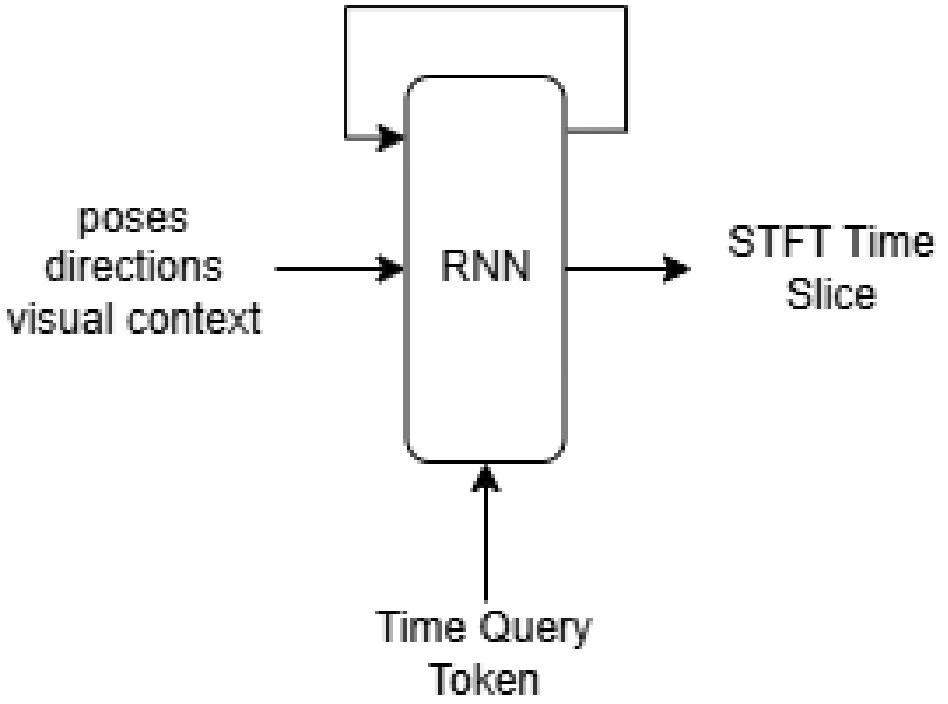
LOW RESOLUTION STFT + UPSAMPLING

- One idea was to predict low band STFT all at once then upscale it one slice at a time.
- Since it is rough and out problems are actually fine frequency detail and fine temporal detail, this doesn't add much, most probably faces another non-identifiability issue
- It might have better temporal modelling of the RIR prior to full prediction, but still I'm suspicious because of it's non-identifiability potential



LSTM AUTOREGRESSIVE

- Autoregressive but using LSTM
- Discarded for old nature of the model

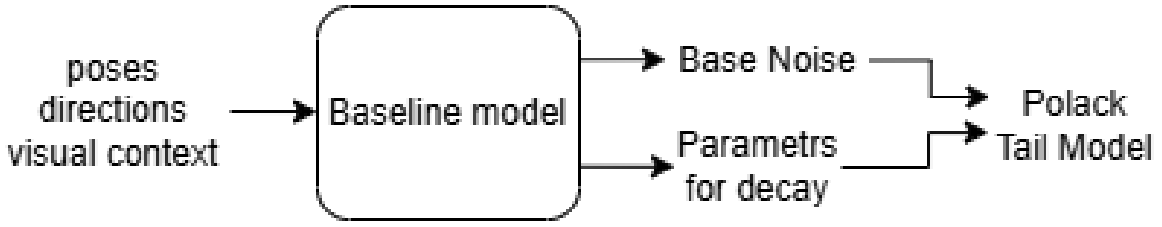


STATEFUL AUTOREGRESSIVE

- The idea is to predict STFT one slice at a time, with previous slices encoded and inputted to model
- Might work but attention does it better, so no point in testing this

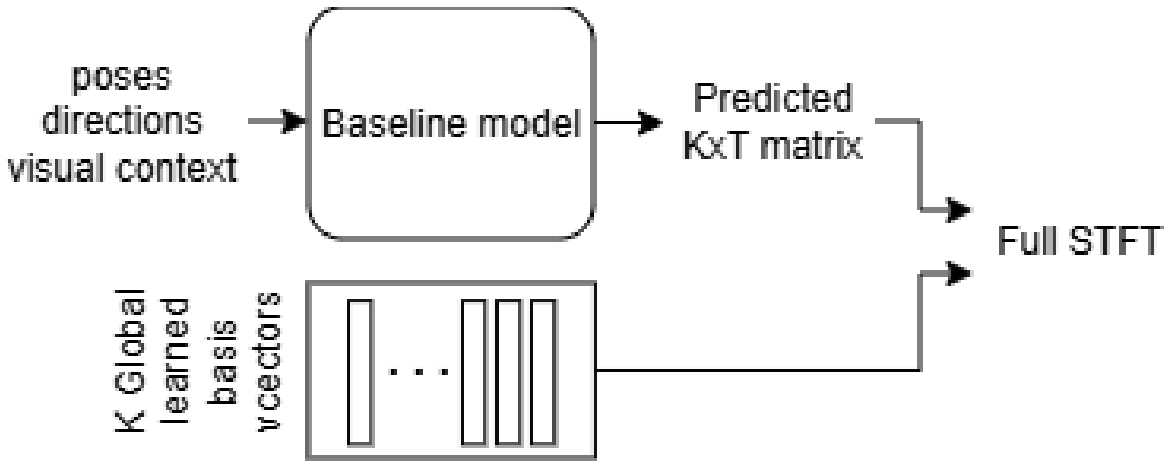
POLACK/FDN PARAMETRIC MODELLING

- Main idea is using a classic model and predicts parameters for classic models (either multi-band Polack or FDN) instead of analytic calculation, then model LR
- Problem is, all these models are just approximations and they most probably will fail in front of full neural models
- Also I tested with how much of STFT they can recreate, at best they can predict STFT with 60% accuracy.



LOW RANK TF FACTORIZATION

- The idea is that the scene has some global modes and by learning them as basis vectors per scene, we can linearly combine them to make any RIR. So instead of predicting a full 513x60 STFT, we just predict combination matrix of these basis vectors
- I tested if the domain is low-rank and it was true, but mostly at lower frequencies (explained in earlier slides)
- This might still be good and useful, but I think full prediction of low frequencies might work better than Non-negative Matrix Factorization (NMF) models
- If multi head prediction for modal head fails, this is the fallback idea



PLAN

TESTS

- Test Attention Auto Regressive (1 week)
- Test dual branch models (1 week)
- Test various loss functions (1 week)