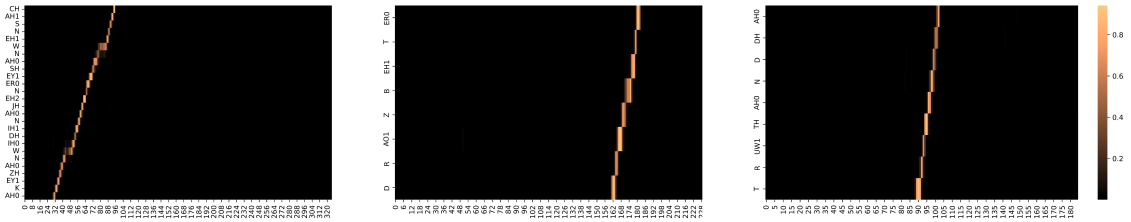
# Supplementary material for DAE-TSE

**Anonymous Submission**

## A Examples of Attention Heatmaps

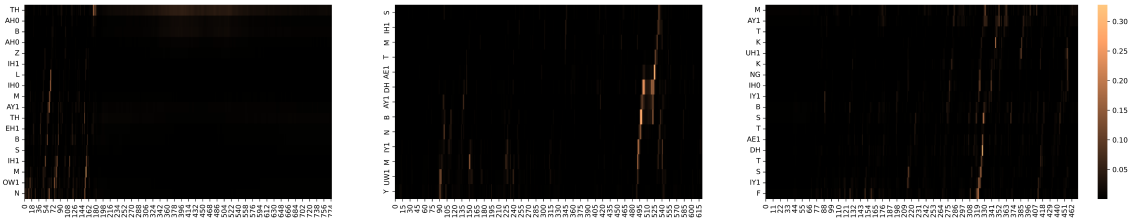
As shown in Figure 1, we present additional examples of cross-attention heatmaps illustrating the relationship between the speech mixture and the keyword. When keywords are present in the mixture, the cue encoder can accurately detect and localize them, producing clear and aligned attention patterns. In contrast, when keywords are absent, the attention map becomes scattered and lacks consistent alignment. These observations further demonstrate the effectiveness of our keyword-guided cue encoder.



(a) Positive sample.
**Transcription**: This was the first **occasion within a generation when such** an entertainment had been given at elmhurst and the only one within the memory of man where the neighbors and country people had been invited guests.
**Keywords**: occasion within a generation when such.

(b) Positive sample.
**Transcription**: She's wonderful more wonderful than anybody we've ever seen anywhere and she **draws better** than the teacher in charlestown.
**Keywords**: draws better.

(c) Positive sample.
**Transcription**: Did not christ himself say I am the way and the **truth and the** life no man cometh unto the father but by me.
**Keywords**: truth and the.



(d) Negative sample.
**Transcription**: From rubbing shoulders with scientists in our little universe by the botanical gardens the boy had come to know a thing or two.
**Keywords**: know miss beth I'm elizabeth.

(e) Negative sample.
**Transcription**: O tis the first tis flattery in my seeing and my great mind most kingly drinks it up mine eye well knows what with his gust is greeing and to his palate doth prepare the cup if it be poison'd tis the lesser sin that mine eye loves it and doth first begin.
**Keywords**: you mean buy that MIS.

(f) Negative sample.
**Transcription**: Totty however had descended from her chair with great swiftness and was already in retreat towards the dairy with a sort of waddling run and an amount of fat on the nape of her neck which made her look like the metamorphosis of a white suckling pig.
**Keywords**: feast that's being cooked I'm.

Figure 1: Cross-attention heatmaps for 3 positive samples (above, keywords present in the mixture) and 3 negative samples (below, keywords absent). The X-axis represents speech frame indices, and the Y-axis corresponds to the phoneme sequence of the keywords.