Tencent AI Lab

Contrastive Separative Coding for Self-supervised Representation Learning

Jun Wang, Max W. Y. Lam, Dan Su, Dong Yu

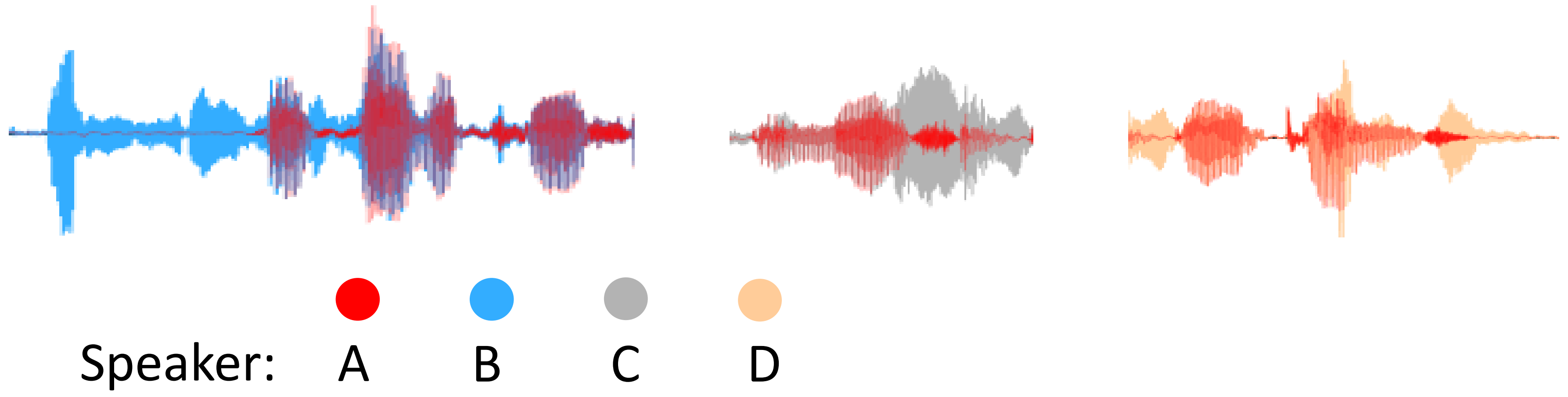
IEEE ICASSP 2021, 2021-04-22

Background and Motivation

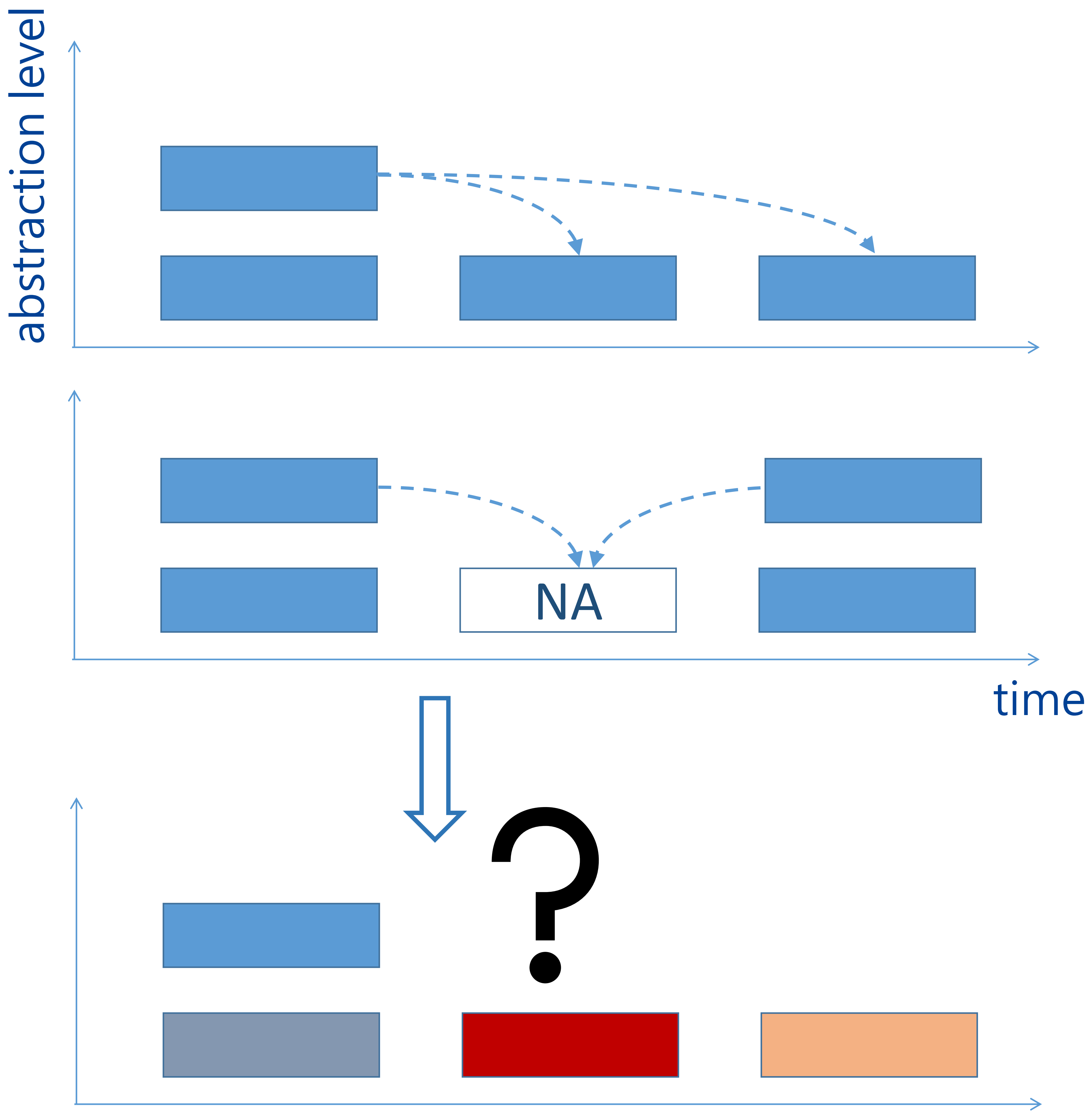
Prior hypothesis:

... predicting the future, neighboring, or missing samples, etc.

The realistic scenarios, however:

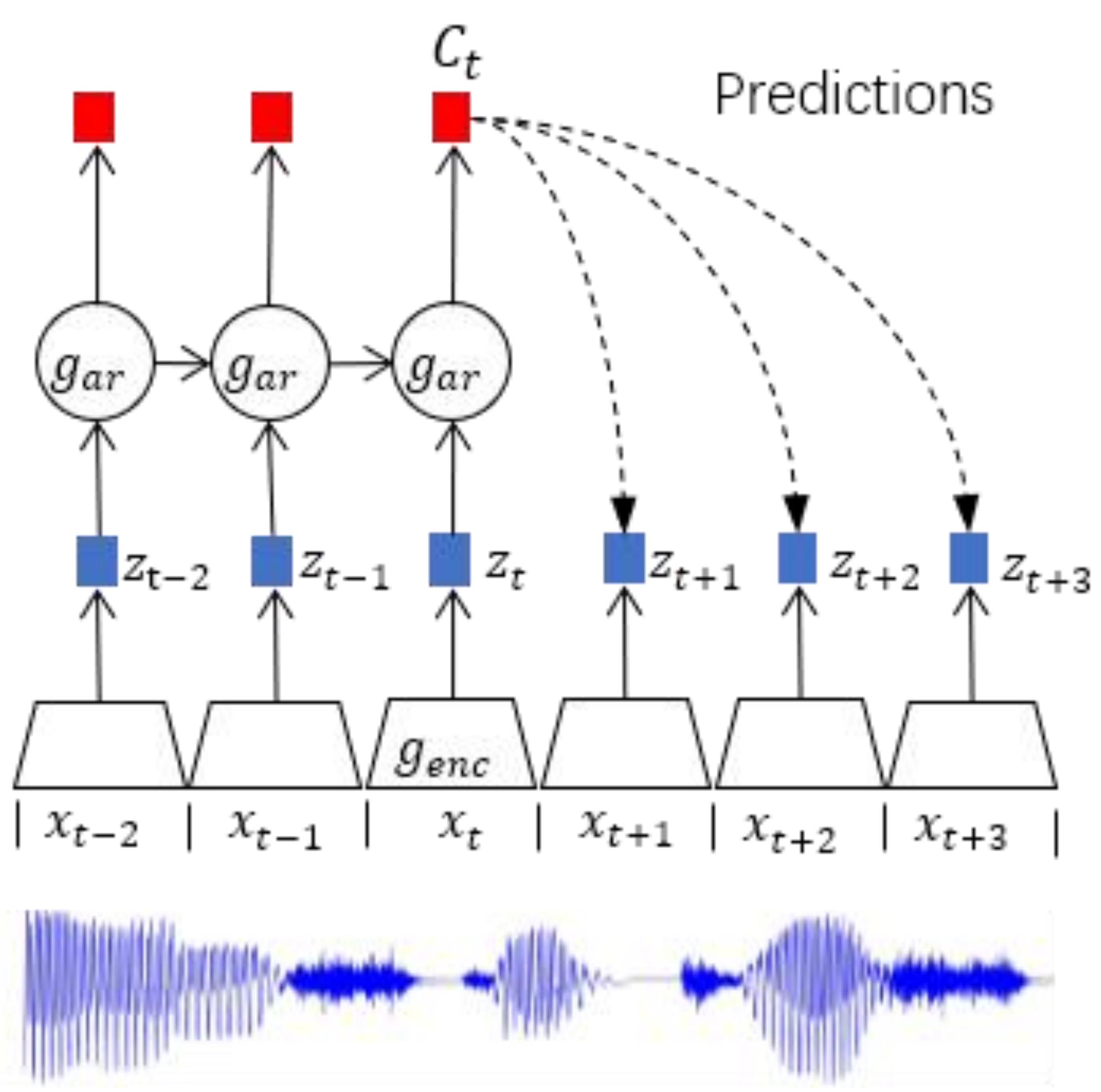


Prior contrastive approaches:

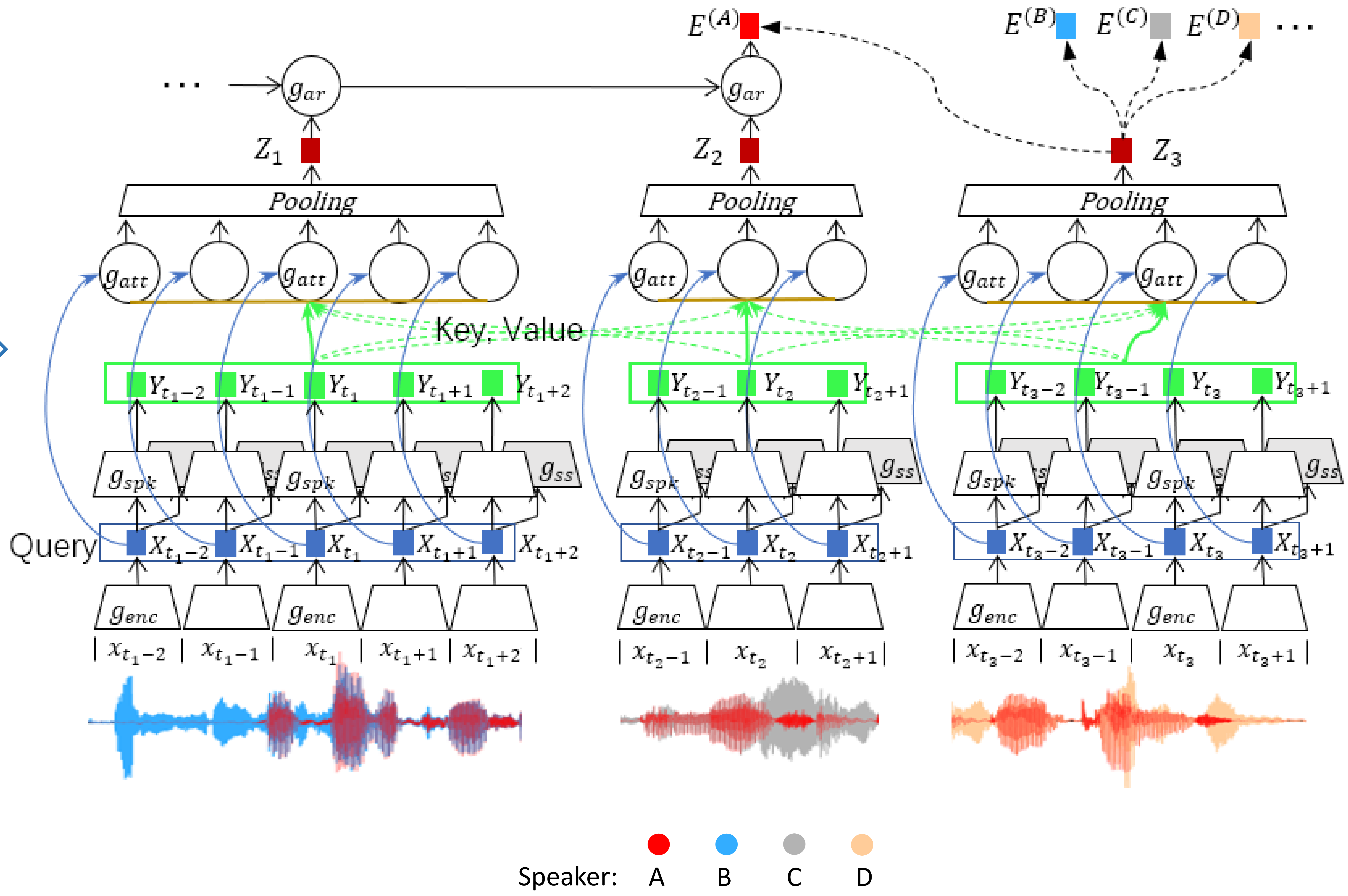


Contrastive Separative Coding

Contrastive Predictive Coding (CPC)

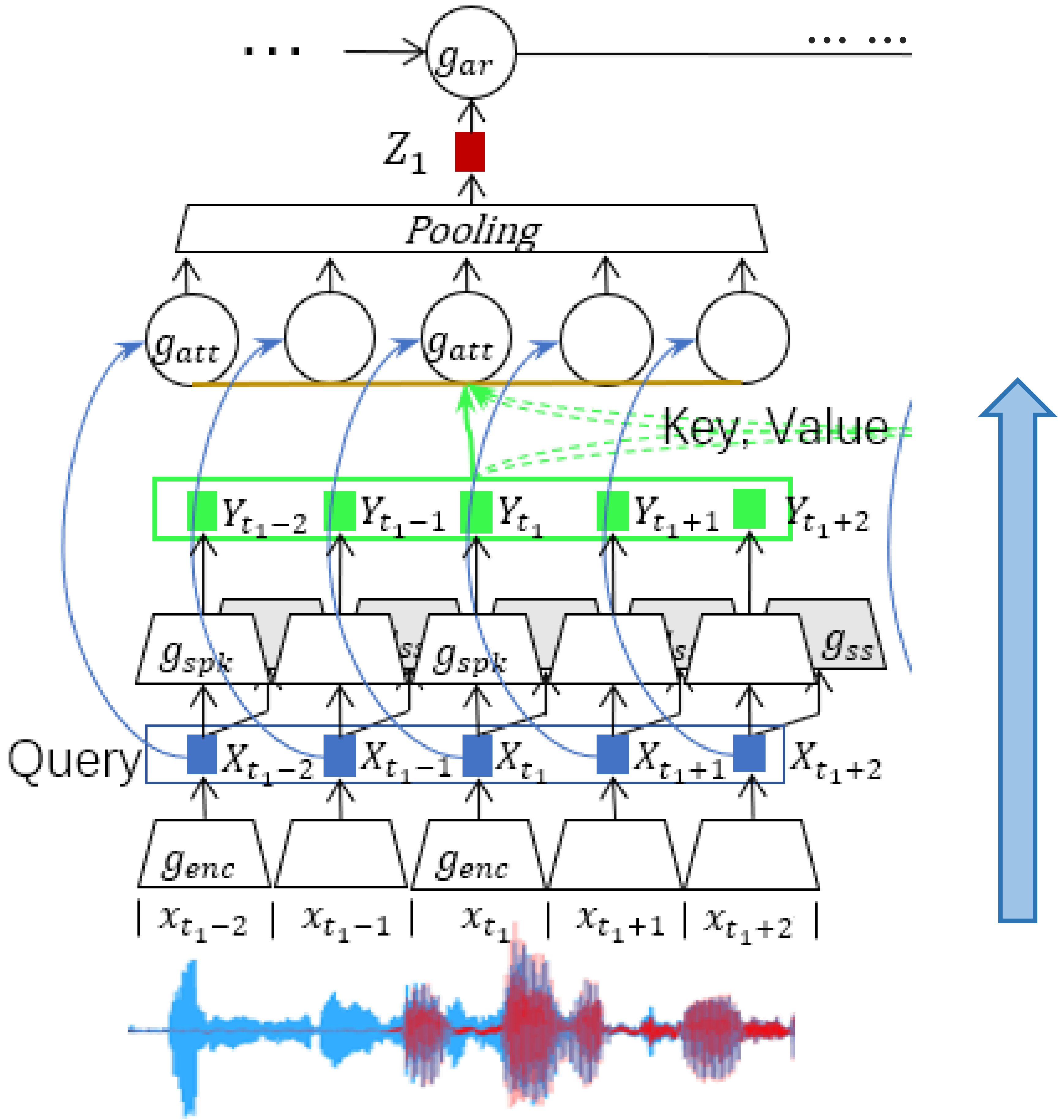


Contrastive Separative Coding (CSC)



Contrastive Separative Coding

CSC



Bottom-up cross attention:

$$\mathbf{a}_{i,j} = \mathrm{softmax}(\mathrm{Query}(\mathbf{X}_i)^\top \cdot \mathrm{Key}(\mathbf{Y}_j)). \tag{1}$$

$$\mathbf{Z}_i = \frac{1}{S_i} \sum_{S_i} \sum_{S_j} \mathbf{a}_{i,j} \cdot \mathrm{Value}(\mathbf{Y}_j)^\top. \tag{2}$$

Contrastive Separative Coding

Tencent AI Lab

CSC Loss:

$$\mathcal{L}_{\mathrm{CSC}} = -\mathbb{E}_{\mathcal{D}}\left[\log\left(f(\mathbf{Z}^{(n_c)}, \mathbf{E}^{(n_c)}) \Big/ \sum_{n=1}^{N} f(\mathbf{Z}^{(n_c)}, \mathbf{E}^{(n)})\right)\right], \tag{3}$$

$$f(\mathbf{Z}^{(n_c)}, \mathbf{E}^{(n)}) = \exp\left(-\alpha\|\mathbf{Z}^{(n_c)} - \mathbf{E}^{(n)}\|_2^2\right), \tag{4}$$

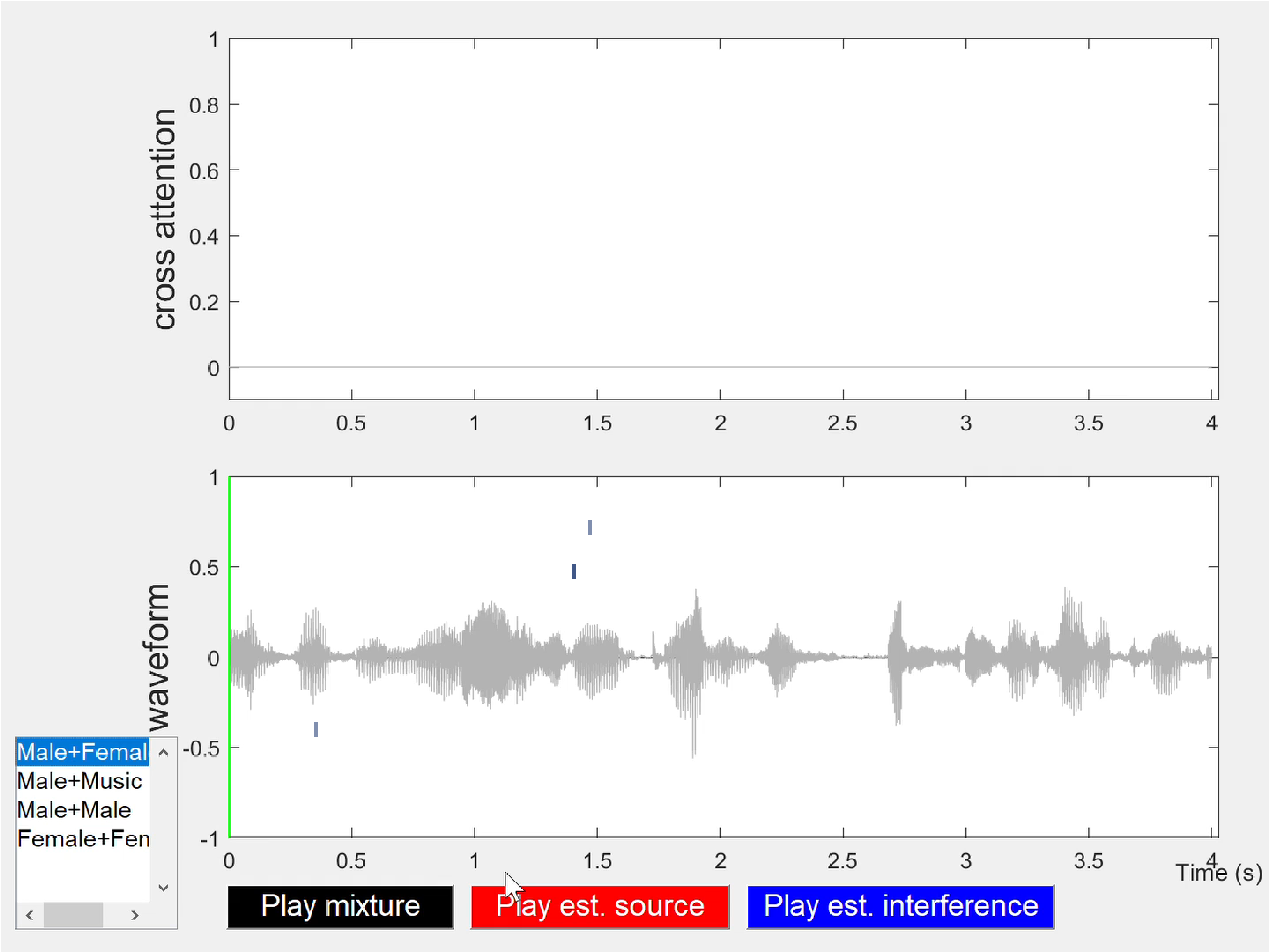
- CSC loss $\mathcal{L}_{\mathrm{CSC}}$ serves as an upper bound of the negative mutual information (MI)
- Minimizing the CSC loss $\mathcal{L}_{\mathrm{CSC}}$ results in maximizing the MI between a global speaker vector and a separative embedding

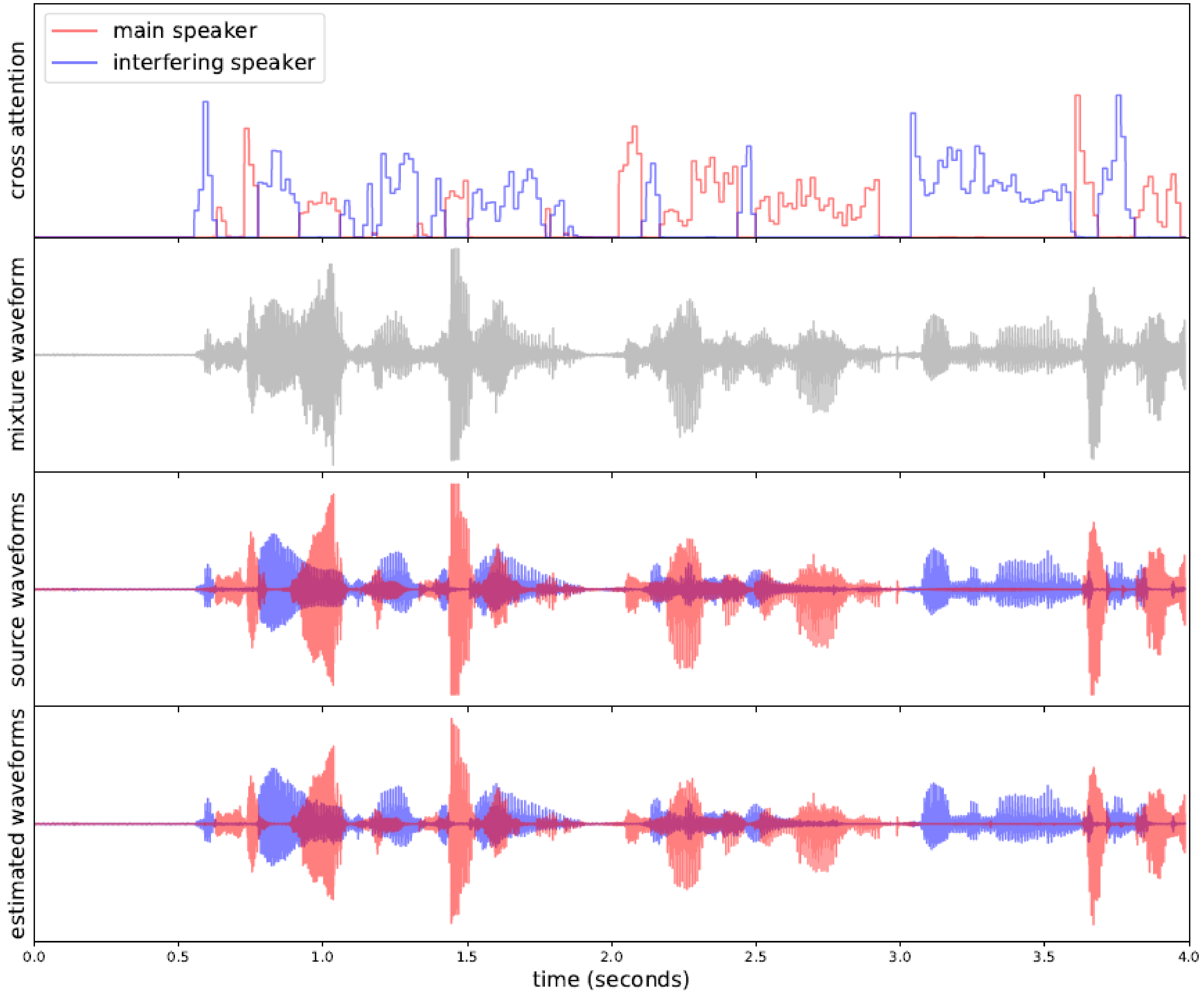
- Applying the proposed $f(\mathbf{Z}, \mathbf{E})$ to $\mathcal{L}_{\mathrm{CSC}}$ corresponds to treating each global speaker vector $\mathbf{E}$ as a cluster centroid (Gaussian mean) of different separative embedding vectors $\mathbf{Z}$ with a learnable parameter $\alpha > 0$ controlling the cluster size (Gaussian variance)
- With our proposed form of $f(\mathbf{Z}, \mathbf{E})$, minimizing $\mathcal{L}_{\mathrm{CSC}}$ results in minimizing the distance between the separative embedding $\mathbf{Z}$ and the corresponding global speaker vector $\mathbf{E}$ meanwhile maximizing the distance between other global speaker vectors
- CSC loss is a rescaled L-2 normalization of *InfoNCE* loss proposed in CPC.

Note: Please see the article for the detailed assumptions and proof for the claims here.

Tencent AI Lab

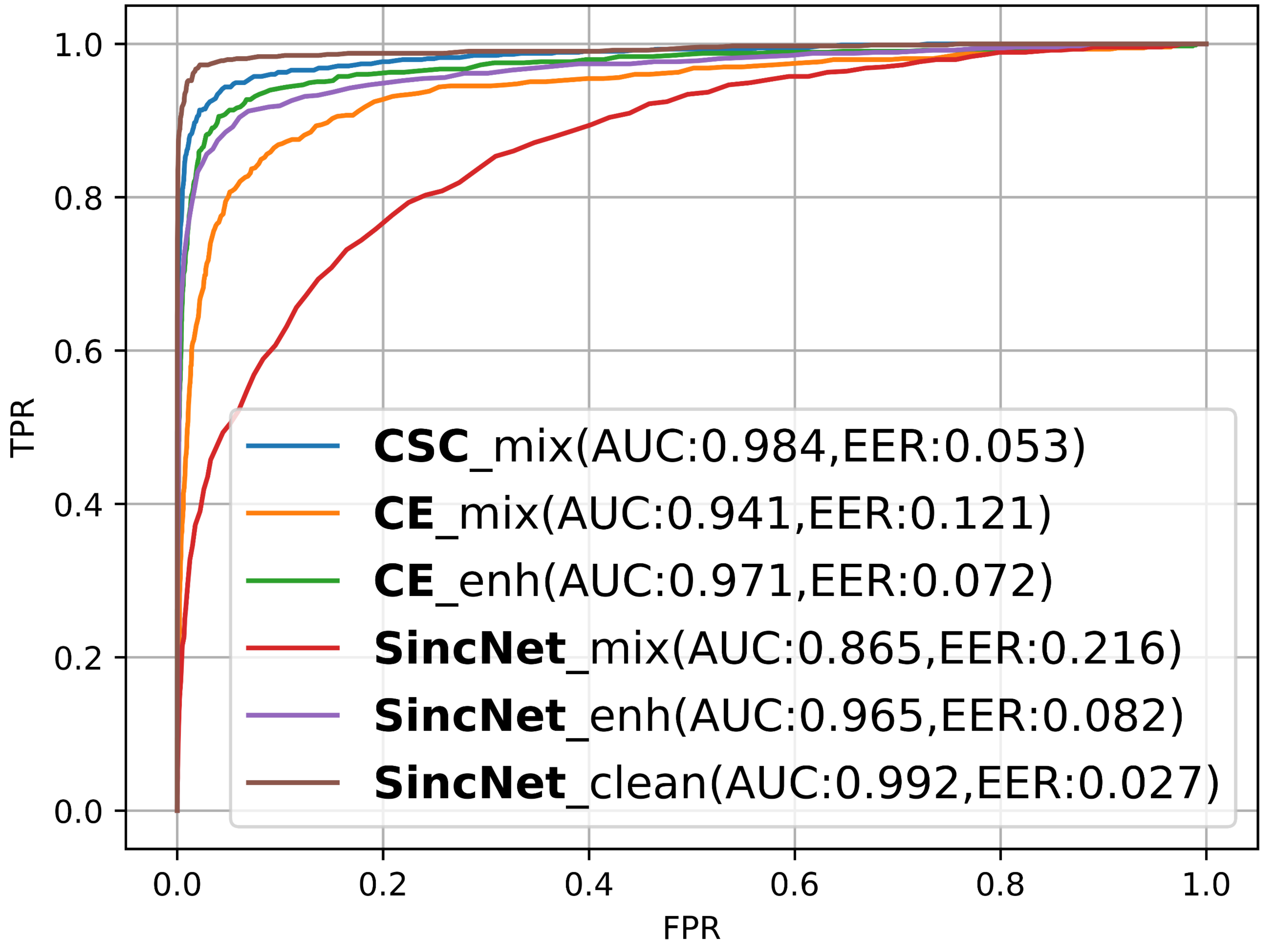
Cross-attention: interpretable mechanism, and improved transparency





Experiments

- Ref. system 1: a conventional speaker-vector-based neural network (SincNet)
- Ref. system 2: ablated the proposed method by removing the g_{att} and g_{ss} models from the graph and replacing the proposed loss with a Cross-Entropy loss (CE)
- Metrics: Equal Error Rate (EER) and Area Under Curve (AUC) on the SV task
- Conditions: "[ ]_mix", "[ ]_enh", and "[ ]_clean": training and test on mixture data, enhanced data by a SOTA SS pre-processing, and clean data
- Results: Our proposed system's performance significantly surpass all reference models', particularly, which in complex interfering conditions is approaching the performance by conventional Ref. 1 in a clean condition.



Conclusions

Tencent AI Lab

- A novel Contrastive Separative Coding (CSC) method is proposed to draw useful representations from complex interfered signals;
- The proposed CSC loss is proved to have in-depth theoretical relations with the mutual information estimation and maximization, as well as prior contrastive learning methods;
- The learned representation have strong discriminability that its complex-condition performance is even approaching the clean-condition performance of a conventional SV system;
- An interpretable bottom-up cross attention mechanism is shown effective in extracting the global aggregation of information across different corrupted observations in various interfering conditions, which is interestingly similar to a human' s auditory selective attention, and to be explored on speaker diarization in our future work.

Tencent
AI Lab

THANK YOU