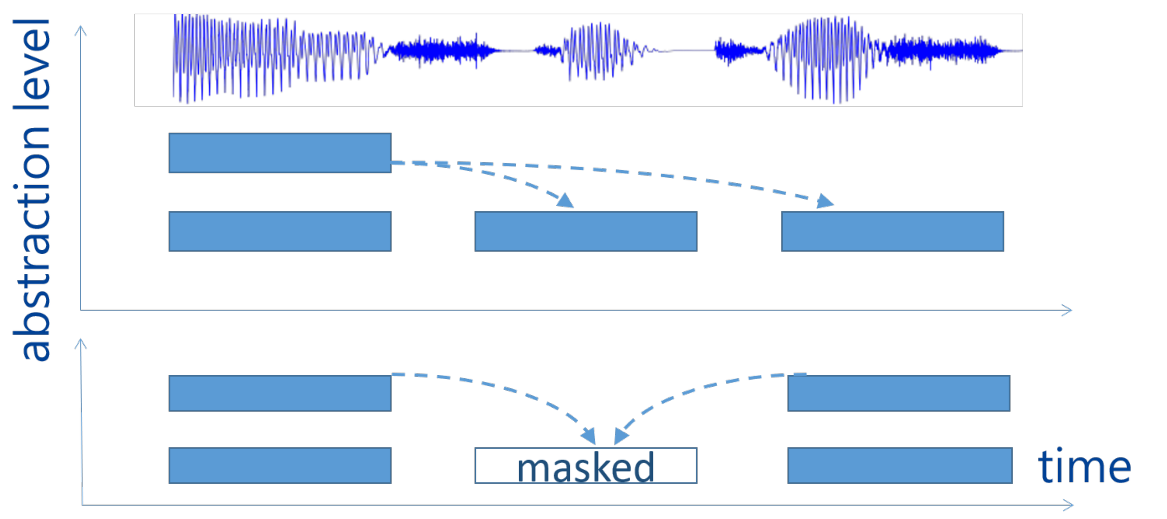
Contrastive Separative Coding for Self-supervised Representation Learning

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

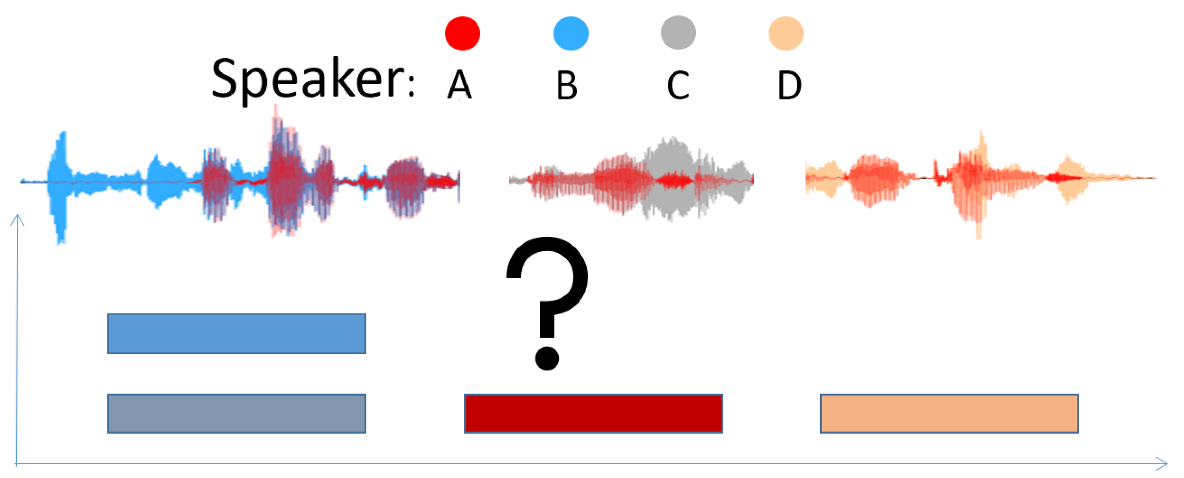
joinerwang@tencent.com, Tencent AI Lab

Introduction

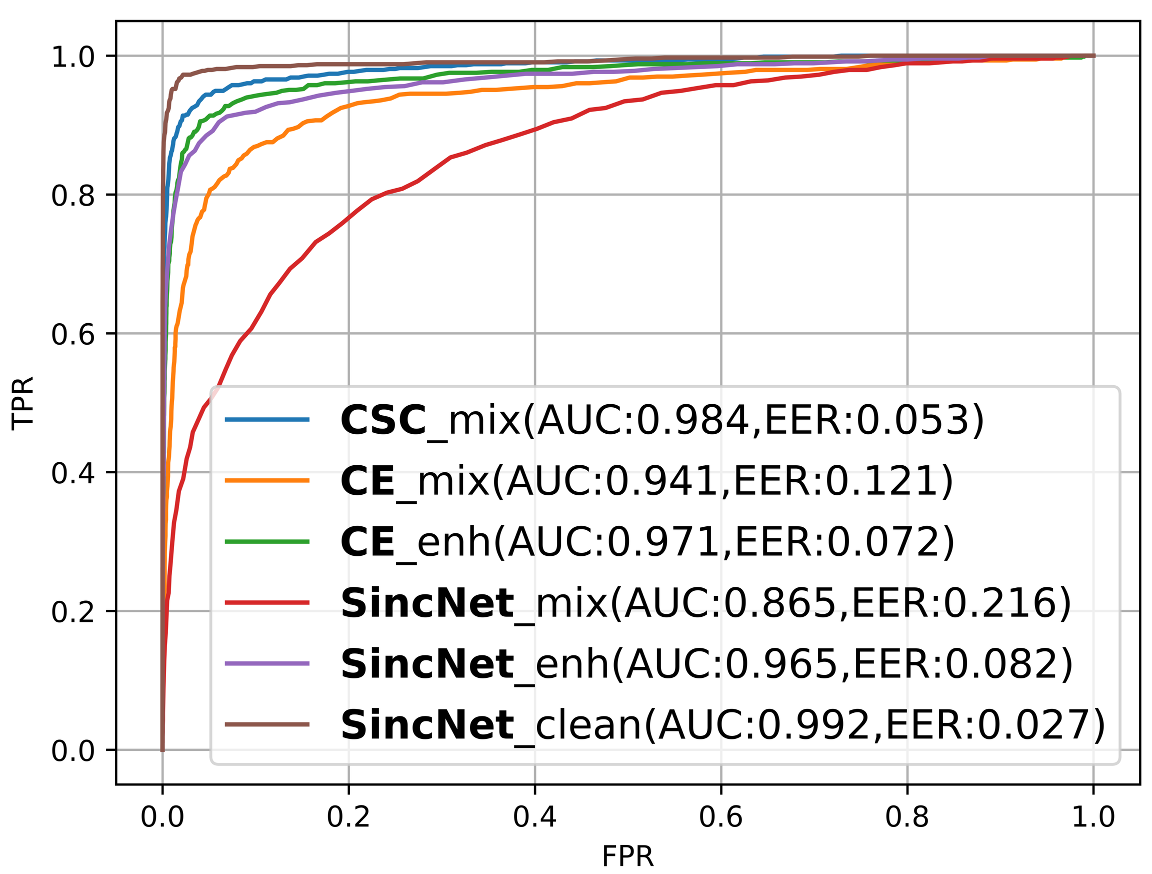
Existing contrastive learning approaches predict the neighboring, missing, or future samples, etc.



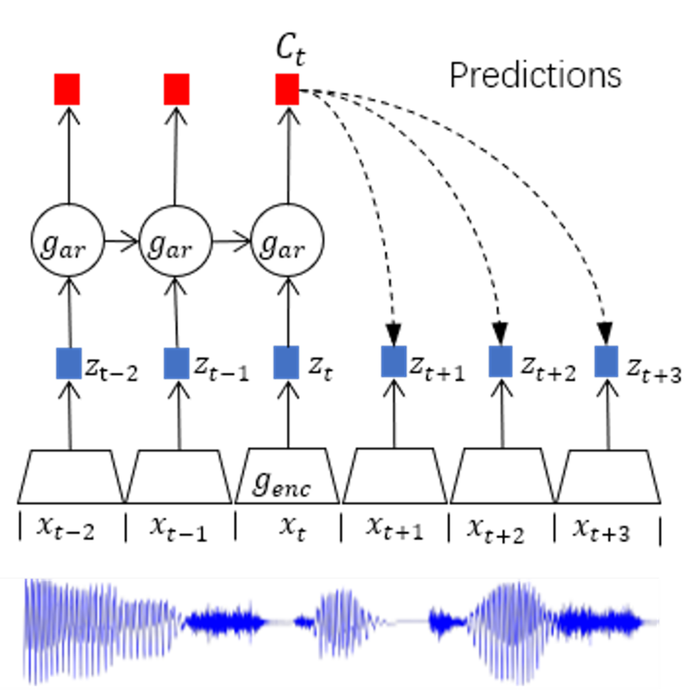
The realistic scenarios, however, have corrupted signals in various interfering conditions; traditionally requiring complicated pipelines to tackle the interference and overlapping segments.



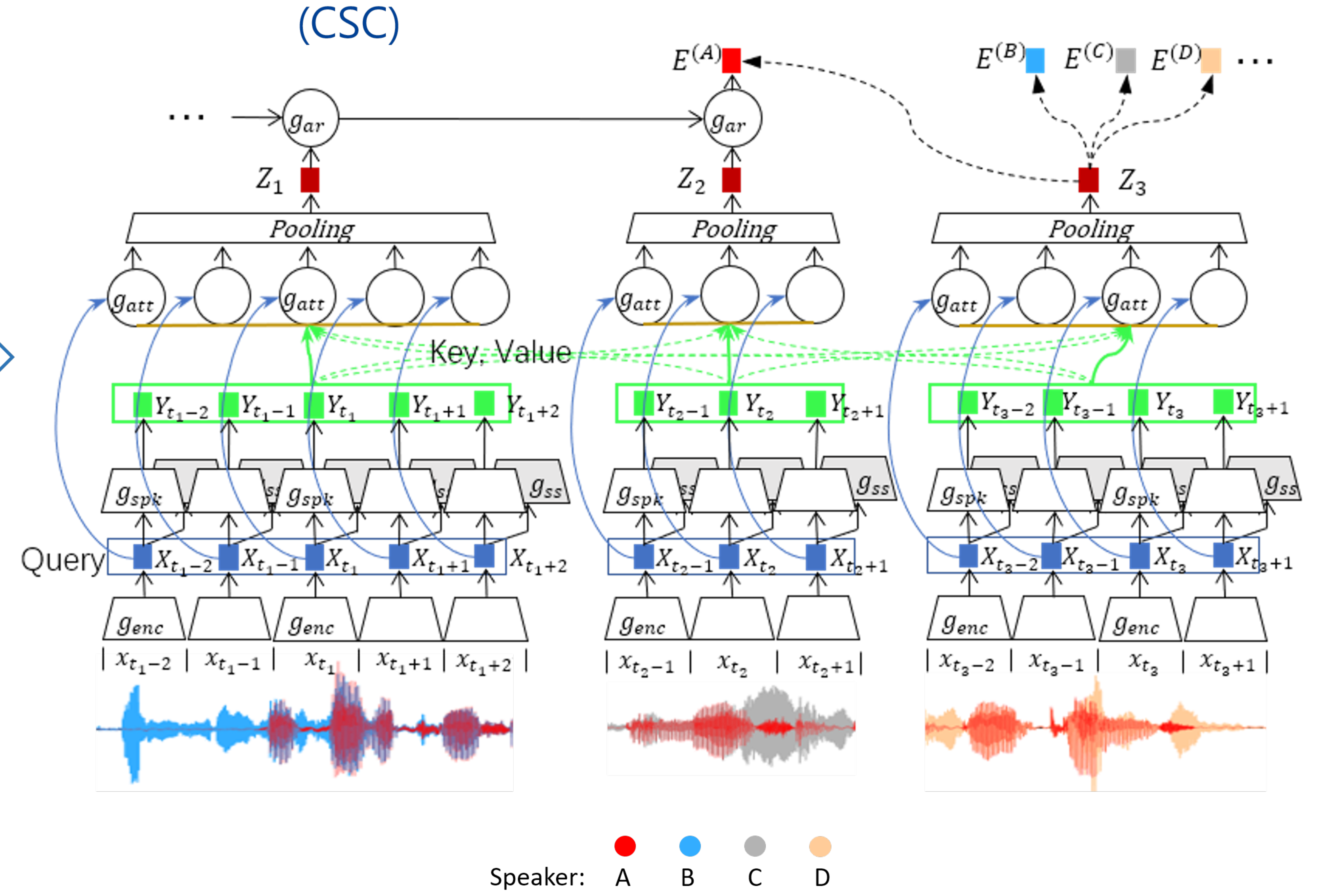
Comparison of SV performances between our proposed method and conventional methods:



Contrastive Predictive Coding (CPC)



Contrastive Separative Coding (CSC)



Proposed Method

Bottom-up cross attention:

- Bottom-up queries from the corrupted signal to retrieve the most relevant representation for the speaker and filter out non-salient, noisy, or redundant parts

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

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

Contrastive Separative Coding (CSC) loss:

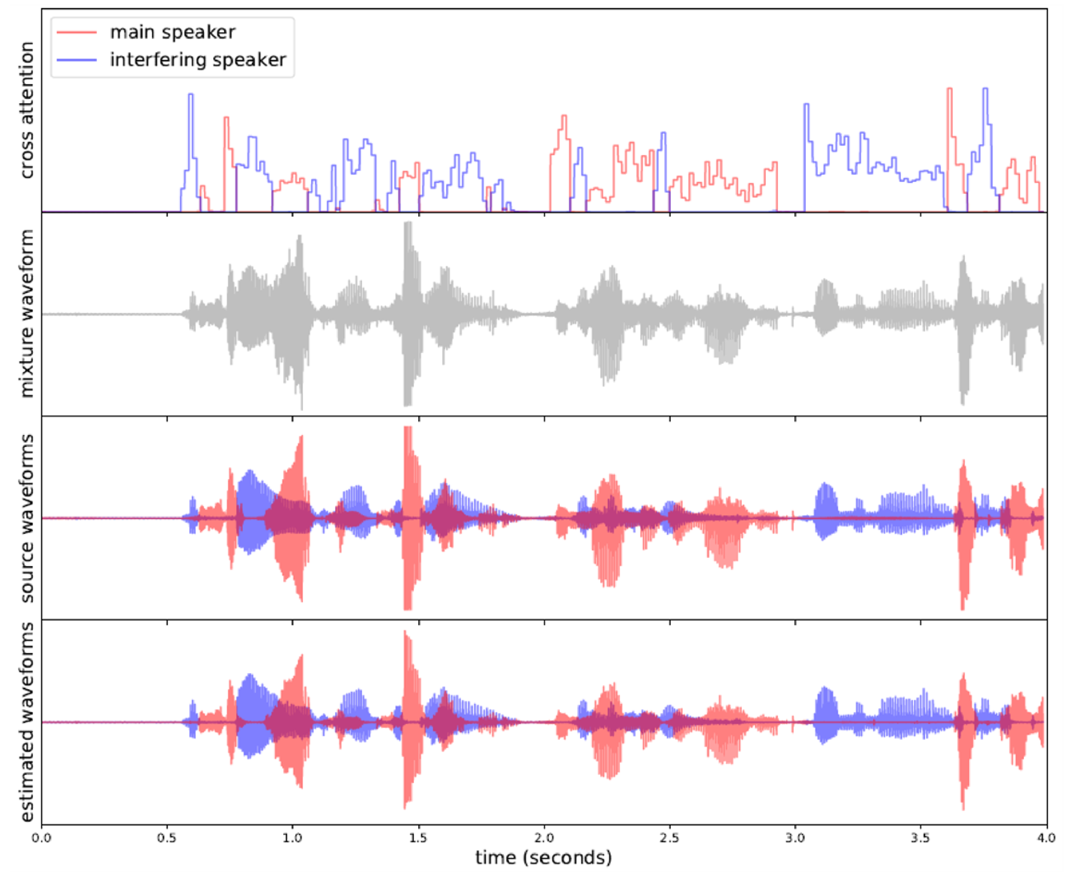
- **CSC** loss $\mathcal{L}_{\text{CSC}}$ serves as an upper bound of the negative mutual information (MI), therefore minimizing the **CSC** loss results in maximizing the MI between a global speaker vector and a separative embedding

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

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

Relation to Prior Art

- Applying the proposed $f(\mathbf{Z}, \mathbf{E})$ 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 $\mathcal{L}_{\text{CSC}}$ 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**.



Result

- Baselines: 1) a conventional speaker-vector-based SV system (**SincNet**), 2): ablation by replacing **CSC** with **CE**
- Conditions: mixture ("[ ]_mix"), enhanced data by a SS pre-processing ("[ ]_enh"), and clean data ("[ ]_clean")
- Results: Ours significantly outperforms the baselines, particularly, ours in complex interfering conditions is approaching the performance by conventional **SincNet** in a clean condition.

Conclusion

- The proposed **CSC** loss is proved to have in-depth theoretical relations with **MI** and **CPC**
- The learned representation can achieve high performances even in very complex conditions
- An interpretable bottom-up cross attention mechanism is shown effective in extracting representations across different observations in various interfering conditions, interestingly similar to an auditory selective attention, to be explored on speaker diarization.