Learning-based Personal Speech Enhancement for Teleconferencing by Exploiting Spatial-Spectral Features

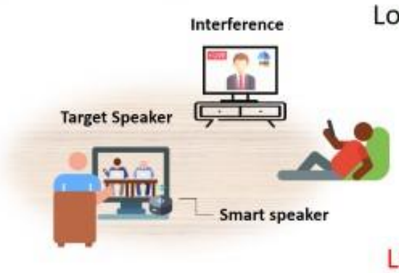
Yicheng Hsu, Yonghan Lee, and Mingsian R. Bai

國立清華大學 NATIONAL TSING HUA UNIVERSITY

icassp 2022 Singapore

Introduction

- Teleconferencing is becoming essential during the COVID-19 pandemic. However, in real-world applications, speech quality can deteriorate due to, for example, background interference, noise, or reverberation.



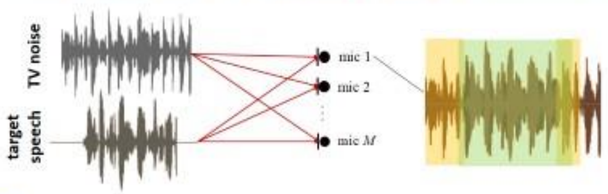
- Target speech extraction from the mixture signals can be performed with the aid of the user's vocal features, including speaker embeddings [1] derived from user enrollment and a novel long-short-term spatial coherence (LSTSC) feature pertaining to the target speaker activity.

Spatial feature extraction

The signal captured by the *m*th microphone can be written as:

$$Y^m(l,f)=\sum_{j=1}^{J}A_j^m(f)S_j(l,f)+V^m(l,f)$$

Acoustic transfer function | Signal of the *j*th source | Nondirectional noise



Target speech absence period: $Y^m(l,f)=H_n^m(f)S_n(l,f)+V^m(l,f)$

Target speech presence period:

$$Y^m(l,f)=H_s^m(f)S_s(l,f)+H_n^m(f)S_n(l,f)+V^m(l,f)$$

Short-term relative transfer function (RTF)

$$\tilde{R}^m(l,f)\equiv\frac{\hat{\Phi}_{y^m y^1}}{\hat{\Phi}_{y^1 y^1}}\equiv\frac{\sum_{n=l-R/2}^{l+R/2}Y^m(n,f)Y^{1*}(n,f)}{\sum_{n=l-R/2}^{l+R/2}Y^1(n,f)Y^{1*}(n,f)}$$

Long-Short-Term Spatial Coherence (LSTSC)

Whitened RTF:

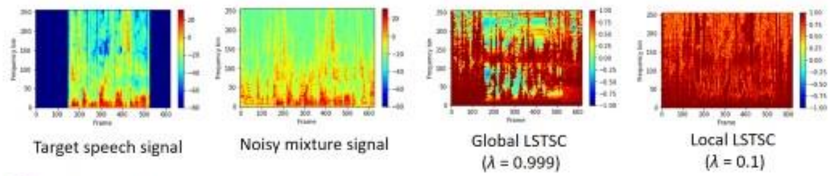
$$\mathbf{r}(l,f)=\left[\frac{\tilde{R}^2(l,f)}{\left|\tilde{R}^2(l,f)\right|},\ldots,\frac{\tilde{R}^M(l,f)}{\left|\tilde{R}^M(l,f)\right|}\right]^T \qquad M:\text{number of microphone}$$

Long-term RTF:

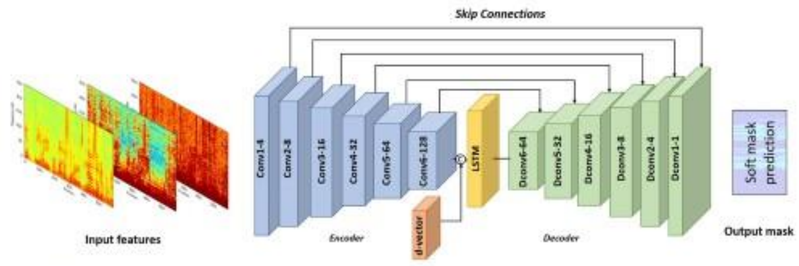
$$\bar{r}^m(l,f)=\lambda\bar{r}^m(l-1,f)+(1-\lambda)r^m(l,f),\ m=2,\ldots,M$$

$$\bar{\mathbf{r}}(l,f)=\left[\frac{\bar{r}^2(l,f)}{\left|\bar{r}^2(l,f)\right|},\ldots,\frac{\bar{r}^M(l,f)}{\left|\bar{r}^M(l,f)\right|}\right]^T \qquad \lambda:\text{forgetting factor}$$

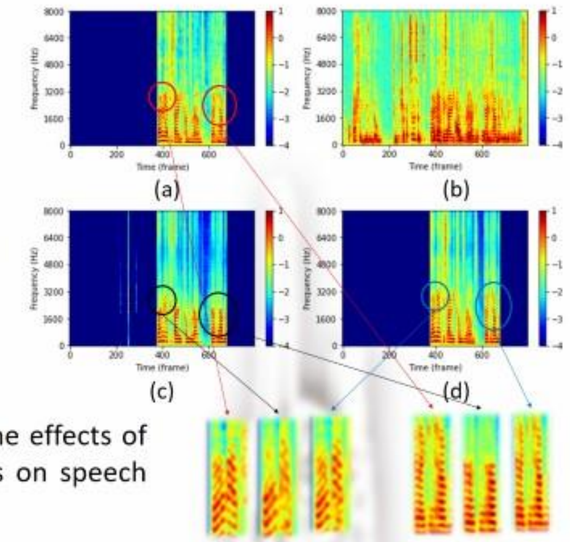
LSTSC: $$\gamma(l,f)\approx\frac{1}{M-1}\sum_{m=2}^{M}\frac{\mathrm{Re}\left\{\tilde{R}^m(l,f)\bar{r}^m(l,f)^*\right\}}{\left|\tilde{R}^m(l,f)\right|\left|\bar{r}^m(l,f)\right|}\approx\frac{1}{M-1}\mathrm{Re}\left\{\mathbf{r}^H(l,f)\bar{\mathbf{r}}(l,f)\right\}$$



Target speech sifting network



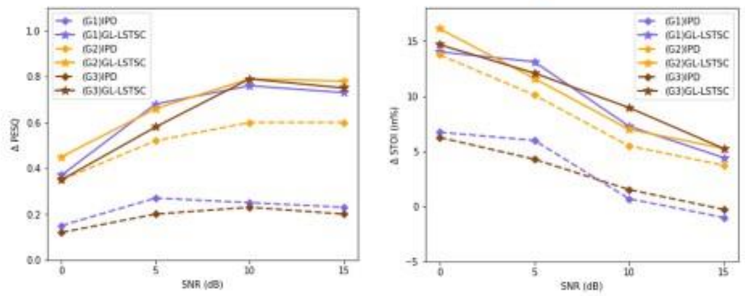
An example of noisy, IPD, the proposed method and clean spectrogram for a speech segment from the test set. (a) clean speech, (b) noisy mixture signal, (c) IPD, and (d) the proposed method (λ = 0.999 and λ = 0.1)

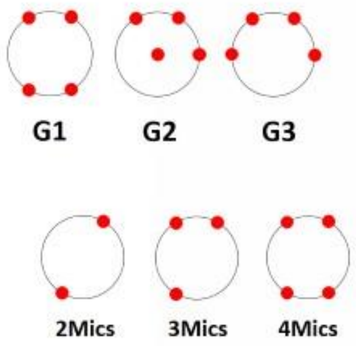


Results

- First, we compared the improvements in performance yielded by the LSTSC feature versus baseline IPD feature.

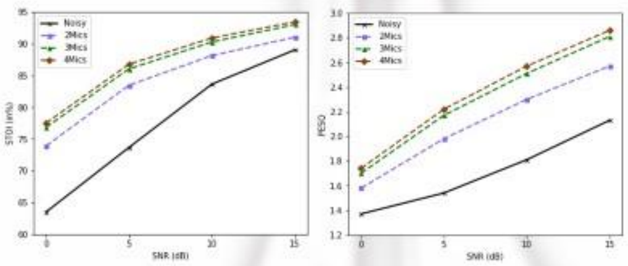
■ STOI and PESQ scores for different array geometries.





- Second, we examined the effects of number of microphones on speech enhancement.

■ STOI and PESQ scores for different number of microphones.



Conclusions

- A **hybrid** TSE system integrating **array signal processing (ASP)** and **deep learning (DL)** has been proposed for teleconferencing applications in the stay-at-home scenarios.
- The spatial feature (LSTSC) is conducive to superior enhancement performance of target speaker and robustness **to unseen RIRs, unseen array geometries**, and **number of microphones**, which is highly desirable in real-world application.

[1] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *Proc IEEE ICASSP*, Calgary, Canada, 2018, pp. 4879–4883.