University Defence Research Centre (UDRC)
In Signal Processing

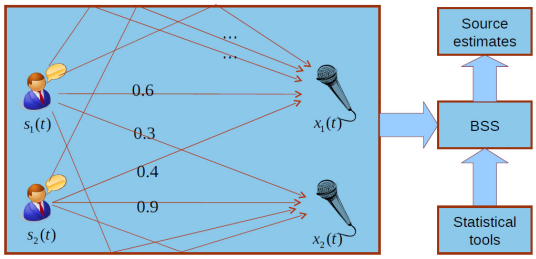
Sponsored by the UK MOD

[O11] Multimodal Blind Source Separation for Robot Audition
Theme: Classification & Multi-modal Processing

PI: Wenwu Wang, University of Surrey
Researchers: W. Wang, P. Jackson, Q. Liu

Status Quo

❑ Blind separation of convolutive speech mixtures



❖ Audio source separation algorithms are usually limited when strong acoustic noise are present in the background.
❖ Visual information from lip-reading helps to isolate speech sources from a mixture of sounds including interfering sounds and background noise.

❑ Aim and Objectives

To use both audio and visual modalities to address the problem of separating target speech signals from multiple competing speech interferences in a room environment for robot applications.

❑ Existing approaches and shortcomings

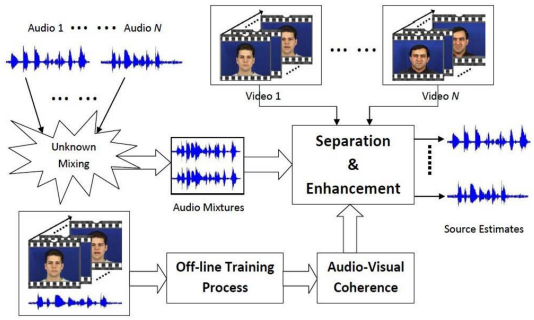
❖ The audio and visual modalities are usually fused on the feature spaces, which need to be synchronised due to the different sampling rates of the audio-visual measurements, based on statistical training using e.g. Gaussian mixture models.
❖Traditional statistical tools either suffer from the over-fitting problem or fail to accurately model the audio-visual coherence.

❑ Anticipated benefits

❖ A robust feature selection scheme is proposed to reduce the over-fitting problem of the statistical training method.
❖ A bimodal dictionary learning method is also developed as an alternative to the statistical method for learning the audio-visual coherence.

Technical Work

❑ System diagram

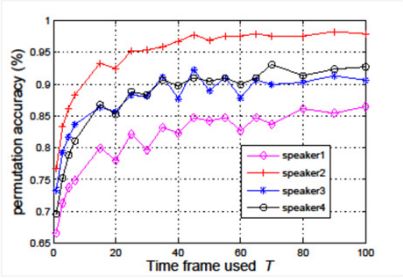


❖ Off-line training process
  ➢ Extraction, synchronisation and fusion of the audio-visual features.
  ➢ Statistical training of audio-visual coherence.
❖ Separation process
  ➢ Frequency domain blind source separation of audio mixtures.
  ➢ Adjust the BSS-separated components via maximizing the statistical audio-visual coherence, or via the visually constrained time-frequency mask generated from dictionary learning.
  ➢ Time domain reconstruction of the speech sources.
❖ Further enhancement
  ➢ Enhancement of the separated sources using spectral subtraction, where the noise power is estimated from the silence period detected via a visual voice activity detection algorithm.

❑ Experimental results

❖ Evaluation of the audio-visual fusion method for addressing the permutation problem of frequency domain blind source separation on the XM2VTS database.

The permutation problem is addressed by bimodal coherence maximization with majority voting.



❖ Signal to interference ratio (SIR in dB) comparison at different signal to noise ratios (SNRs in dB) on the XM2VTS database.

Evaluation of Permutation Indeterminacy Cancellation						
Input SNR		10	15	20	25	30
Input SIR		-1.42	-0.91	-0.74	-0.68	-0.66
Output SIR	before sorting	4.02	6.40	8.81	13.41	13.24
	after sorting	5.29	9.28	12.97	14.66	14.8
Evaluation of Scaling Indeterminacy Cancellation						
Input SNR		4	6	8	10	12
Input SIR		-3.13	-2.38	-1.82	-1.42	-1.15
Output SIR	before scaling	0.66	2.07	3.83	5.29	6.77
	after scaling	2.07	3.71	4.55	5.82	6.80

❑ Summary

❖ Video can be used to enhance audio source separation.
❖ The proposed audio-visual fusion technique mitigates the influence of outliers and improves the performance of the blind source separation algorithm.

Exploitation & Military Relevance

❖ Potential applications include: ***battlefield monitoring***, ***security surveillance***, ***life rescue*** from war-field, ***incident detection*** for public safety, ***human-robot interaction***, ***human-computer interfaces***, ***elderly healthcare***, with possible additional applications in, e.g., ***multimedia products***.

[dstl]

MINISTRY OF DEFENCE

EPSRC
Engineering and Physical Sciences Research Council

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) and the MOD University Defence Research Centre on Signal Processing (UDRC).