Integrating binaural cues and blind source separation method for separating reverberant speech mixtures

Atiyeh Alinaghi, Wenwu Wang and Philip JB Jackson

Centre for Vision, Speech and Signal Processing (CVSSP), Department of Electronic Engineering (FEPS), University of Surrey, Guildford GU2 7XH, UK

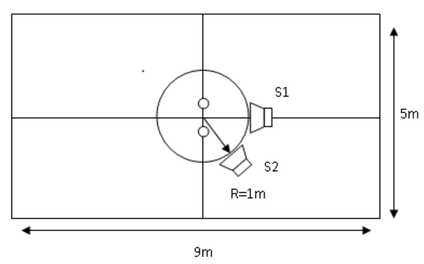
a.alinaghi@surrey.ac.uk

INTRODUCTION

Unlike humans, machines like Automatic Speech Recognition (ASR) systems are not very efficient in recognizing and localizing a specific sound source in the presence of other sources.

Methods to improve their efficiency:
- Blind Source Separation (BSS); based on the statistical properties of the sources
- Computational Auditory Scene Analysis (CASA); inspired by the human auditory system

This work combines the two approaches.



METHODS

• Model-based Expectation Maximization Source Separation and Localization (MESSL)

Estimates a soft mask to segregate the source signals by exploiting the EM algorithm to iteratively refine the time-frequency (T-F) regions allocated to one specific source:
- E-step: the time-frequency (T-F) points which fit a specific source model are assigned to that source.
- M-step: each source model's parameters get updated based on the T-F points assigned to that source.

Interaural Level Differences (ILDs) and Interaural Phase Differences (IPDs) are used to model each source by Gaussian distributions and the parameters are estimated using maximum likelihood.

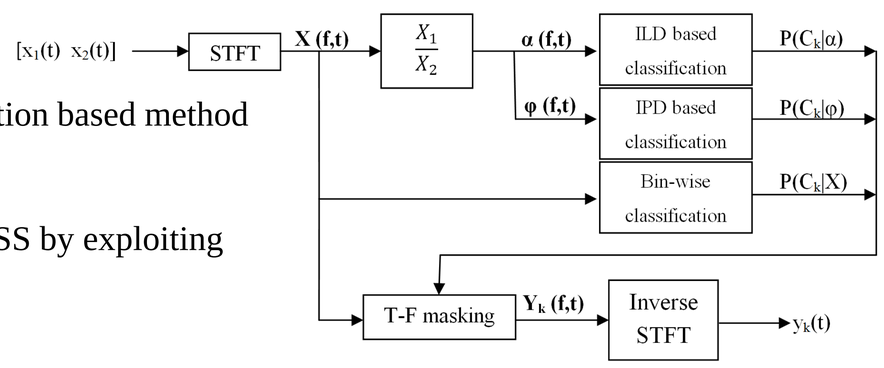
•Underdetermined Blind Source Separation for convolutive mixtures (BSS)

In this method the two measured signals are put together to form a new data whose elements are 2-D vectors. Then the T-F points are grouped into N clusters (N: the number of sources). The probability of each T-F point belonging to one specific source is also calculated similar to MESSL algorithm.

• Combination of the two methods:

○ To improve the performance of MESSL Which a localization based method for closely spaced sources

○To solve the permutation problem of frequency domain BSS by exploiting the MESSL approach for reliable initialization



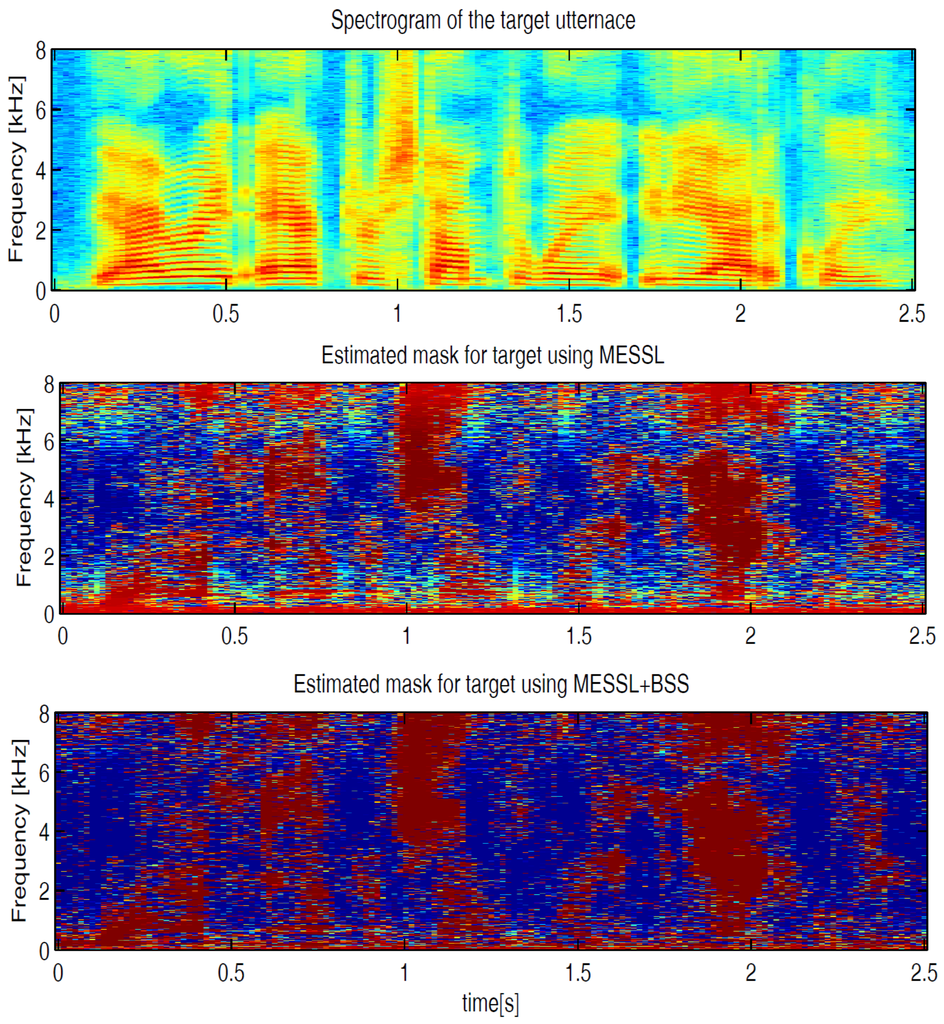
$$L(\hat{\Theta}) = \max_{\theta} \sum_{\omega,t} \log p(\phi(\omega,t;\tau), \alpha(\omega,t), \mathbf{x}(\omega,t)|\Theta)$$

$$\hat{\Theta} = \{\xi_{i,\tau}(\omega), \sigma_{i,\tau}(\omega), \mu_i(\omega), \eta_i(\omega), \mathbf{a}_k(\omega), \gamma_k(\omega), \psi_{i,\tau}(\omega)\}$$

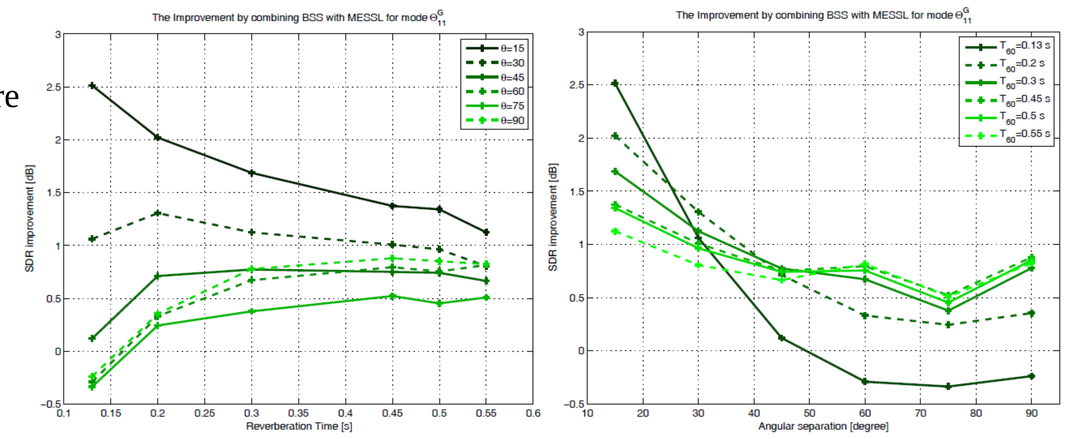
EXPERIMENTS & RESULTS

- Mixtures generated by adding the interfering signals at θ ={15º ,30º,..., 90º} to the target signals at θ =0º with different reverberation times T60={0.13 s, 0.2 s,0.3 s, 0.45 s, 0.5 s, 0.55 s}
- MESSL algorithm applied to 15 mixtures for each θ and T60 with various complexity represented by different modes
- Combined method also applied to the same mixtures to compare the algorithms
- Method performances evaluated by signal-to-distortion ratio (SDR) of the segregated signal.

modes	T60=0.45 s MESSL							modes	MESSL with BSS						
$\Theta_{ild,ipd}$	15º	30º	45º	60º	75º	90º	Ave	$\Theta_{ild,ipd}$	15º	30º	45º	60º	75º	90º	Ave
Θ^G_{00}	1.56	2.61	2.82	2.26	3.42	1.46	2.35	Θ^G_{00}	3.08	3.75	3.69	3.20	3.99	2.45	3.36
Θ^G_{01}	1.41	2.44	2.63	2.10	3.19	1.39	2.19	Θ^G_{01}	3.17	3.80	3.74	3.30	4.03	2.60	3.44
Θ^G_{10}	2.02	3.00	3.22	2.76	3.67	1.92	2.77	Θ^G_{10}	3.20	3.83	3.84	3.44	4.04	2.70	3.51
Θ^G_{11}	1.87	2.85	3.13	2.72	3.54	1.95	2.68	Θ^G_{11}	**3.24**	**3.86**	**3.88**	**3.52**	**4.06**	**2.83**	**3.56**
$\Theta^G_{0\Omega}$	2.05	2.92	2.97	2.38	3.46	1.64	2.57	$\Theta^G_{0\Omega}$	2.84	3.65	3.60	3.05	3.85	2.43	3.24
$\Theta^G_{\Omega 0}$	**2.34**	**3.10**	3.29	2.77	**3.69**	1.94	2.85	$\Theta^G_{\Omega 0}$	2.80	3.42	3.57	3.13	3.84	2.42	3.20
$\Theta^G_{\Omega 1}$	2.19	3.01	3.26	2.76	3.58	1.97	2.79	$\Theta^G_{\Omega 1}$	2.79	3.49	3.62	3.16	3.84	2.48	3.23
$\Theta^G_{1\Omega}$	2.04	2.77	3.22	2.81	3.56	2.05	2.74	$\Theta^G_{1\Omega}$	2.62	3.34	3.64	3.27	3.84	2.62	3.22
$\Theta^G_{\Omega\Omega}$	2.34	2.96	**3.34**	**2.92**	3.66	**2.14**	**2.89**	$\Theta^G_{\Omega\Omega}$	2.69	3.34	3.52	3.20	3.77	2.58	3.18



The pattern in the masks is the same, but more distinct areas are in the combined method showing that each time frequency point is assigned to one source with more confidence



(a) Different angular separation (b) Different T_{60}

CONCLUSIONS

- More improvement for simpler modes compared to complex ones
- More considerable improvement for small separation angles which is highly desirable
- Improvement converges as the reverberation time increases
- Best MESSL performance occurs at most complex mode (frequency dependent) while the combined method works much better at frequency independent mode.