Microphone Array Power Ratio for Speech Quality Assessment in Noisy Reverberant Environments [1]

Prof. Israel Cohen

Department of Electrical Engineering
Technion - Israel Institute of Technology
Technion City, Haifa 3200003, Israel

IWAENC 2016

[1] Joint work with Reuven Berkun (Technion) and Baruch Berdugo (Phoenix Audio Technologies)

Outline

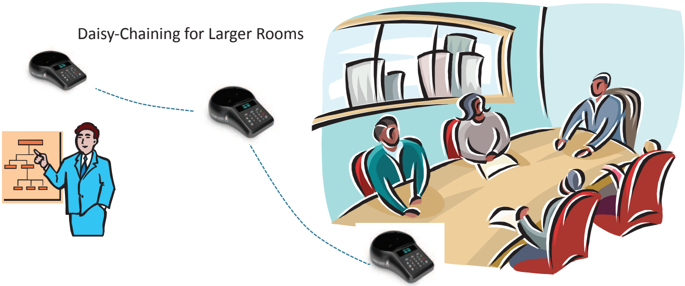
Hands-free communication systems

Enhancement of speech signals is of great interest in many hands-free communication systems:

- Hearing-aids devices.
- Cell phones and hands-free accessories for wireless communication systems.
- Conference and telephone speakerphones.
- Etc.

Teleconferencing

- **Teleconferencing in large rooms:** Use more than one microphone for audio pickup.
- **A major challenge:** Monitor the perceived quality of each microphone signal and select, at any given point in time, the microphone with the best reception.



Teleconferencing (cont.)

- Microphones that are used in industrial applications are generally **not calibrated**.
- The **sensitivities** of different microphones may be quite different.
- Therefore, the **power** is not reliable for a comparison between signals measured with different microphones (Wolf and Nadeu, 2010).
- The **signal-to-noise ratio** is also not a reliable measure to quantify the level of reverberation, since in real applications, the noise cannot be assumed uniform, nor the late reverberation is uniform (Obuchi, 2004, Wölfel et al., 2006).

Problem Formulation

- A source signal measured at point $p_i = (x_i, y_i, z_i)$ $(i = 1, 2, \dots, N)$ is given by

$$r_i(t) = s(t) * h_i(t) + n_i(t) \,.$$

- **Perception of the amount of reverberation** in a given signal is closely related to the **direct-to-reverberation ratio**.
- For evaluating the direct-to-reverberation ratio, the impulse response $h_i(t)$ is split into **early (direct) and late (reverberant) parts**:

$$h_i(t) = h_{i,d}(t) + h_{i,r}(t) \,.$$

Problem Formulation (cont.)

- The direct-to-reverberation ratio is defined as the ratio between the energy of the "**direct path**" (including the early reflections) and the energy of the "**reverberant paths**" (containing only the late reflections).

$$\mathrm{DRR} = \frac{E_d}{E_r} = \frac{\int_0^{T_d} h^2(t)dt}{\int_{T_d}^{\infty} h^2(t)dt}$$

- Our objective is to determine which signal out of the given set of measured signals $\{r_i(t)\,|\,i = 1, 2, \ldots, N\}$ has the **greatest direct-to-reverberation ratio**.
- **Real-time quality monitoring** based on **short segments** of the signals, robust to **differences in sensitivities** of microphones and **environmental conditions**.

Related Works

- Channel selection measures for multi-microphone speech recognition (Wolf and Nadeu, 2014)
  - Microphones are arbitrarily located.
  - Position and orientation of the speaker is unknown.
  - Objective: **Rank the channels** as close as possible to the word error rate (WER) based ranking.
  - **Envelope-variance measure**: The effect of reverberation is observed as a reduction in the dynamic range of the speech intensity envelope (Houtgast and Steeneken, 1985).
  - Channel selection provides **significant recognition improvements** (in some cases, up to 46% compared to randomly selected channel).
  - **A good calibration of all microphones is still required**, which is not a trivial task.

Related Works

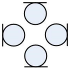
- Acoustic Characterization of Environments (ACE) Challenge (Eaton, Gaubitch, Moore, and Naylor, 2016)
  - The ACE Challenge attracted participation from 9 research teams around the world.
  - Focused on **non-intrusive estimation of the reverberation time (T60) and DRR**.
  - Classes of algorithms:
    1. Analytical with or without bias compensation (ABC);
    2. Single feature with mapping (SFM);
    3. Machine learning with multiple features (MLMF).
  - **Non-intrusive T60 estimation** is a mature field.
  - **Non-intrusive DRR estimation** however is a significantly less mature field: Large biases and MSEs (the best algorithm estimates DRR to within an RMS error of about 3 dB and a $\rho \approx 0.6$ for typical operating scenarios of 1 to 18 dB SNR).

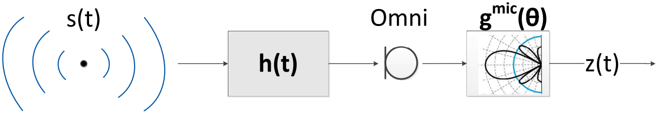
Related Works (cont.)

- **Signal-based** quality measures:
  - Signal-to-diffuse ratio estimation
    - Spatial complex coherence between microphones (Jeub, Nelke, Beaugeant, and Vary, 2011).
    - Direct & diffuse part segregation using beamforming (Thiergart, Ascherl, and Habets, 2014) (Hioka et. al, 2012).
  - Modulation spectral analysis:
    Speech to reverberation modulation energy ratio (SRMR) (Falk, Zheng, and Chan, 2010).
- Generally, **correlation** of signal-based measures with subjective listening tests **is insufficient** (Goetze, Albertin, Kallinger, Mertins, and Kammeyer, 2010).

Configuration

- Unidirectional microphone array
  - Directional elements
  - Beamforming





- $g^{\mathrm{dir/opp}}(\theta)$ - The microphone directional gain at angle θ

Signal Model

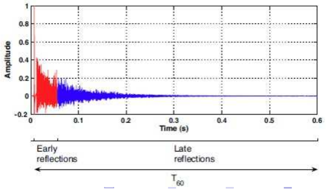
- The measured signal:

$$z(t) = \int_{-\infty}^{t} s(\tau)h(t-\tau)d\tau + v(t),$$

  - $s(t)$ speech signal
  - $h(t)$ room impulse response (RIR)
  - $v(t)$ ambient noise

- Reverberated RIR model:

$$h(t) = \begin{cases} h_d(t), & \text{for } 0 \leq t < T_r \\ h_r(t), & \text{for } t \geq T_r \\ 0, & \text{otherwise,} \end{cases}$$



Signal Model (cont.)

- Statistical room acoustics model (Polack, 1988) (Habets, 2007)
  - $h_d(t) = \begin{cases} b_d(t)e^{-\delta t}, & \text{for } 0 \leq t < T_r \\ 0 & \text{otherwise,} \end{cases}$
  - $b_d(t) \sim \mathcal{N}(0, \sigma_d^2)$
  - $\delta = \frac{3 \ln 10}{T_{60}}$
  - $h_r(t) = \begin{cases} b_r(t)e^{-\delta t}, & \text{for } t \geq T_r \\ 0 & \text{otherwise,} \end{cases}$
  - $b_r(t) \sim \mathcal{N}(0, \sigma_r^2)$

$\Rightarrow$ The measured signal energy:

$$\mathbb{E}_z\{z^2(t)\} = \mathbb{E}_z\{z_d^2(t)\} + \mathbb{E}_z\{z_r^2(t)\}$$

$\lambda_s(t) = \mathbb{E}_s\{s^2(t)\}$, $\mathbb{E}_z\{z_d^2(t)\} = f(\lambda_s(t), \sigma_d^2, T_r)$,
$\mathbb{E}_z\{z_r^2(t)\} = f(\lambda_s(t), \sigma_r^2, T_r)$

Directional Array Response

⇒ The **direct microphone** signal energy:

$$\mathbb{E}_z\{[z^{\text{dir}}(t)]^2\} = [g^{\text{dir}}(\theta)]^2 \cdot \mathbb{E}_z\{z_d^2(t)\} + \frac{1}{\Omega}\int_{\Omega}[g^{\text{dir}}(\theta')]^2 d\theta' \cdot \mathbb{E}_z\{z_r^2(t)\}$$

⇒ The **opposite microphone** signal energy:

$$\mathbb{E}_z\{[z^{\text{opp}}(t)]^2\} = \frac{1}{\Omega}\int_{\Omega}[g^{\text{opp}}(\theta')]^2 d\theta' \cdot \mathbb{E}_z\{z_r^2(t)\}$$

Directional Power Ratio

- Assuming the microphones are calibrated:
  - $\bar{g}^2 = \frac{1}{\Omega}\int_\Omega [g^{\text{dir}}(\theta')]^2 d\theta' = \frac{1}{\Omega}\int_\Omega [g^{\text{opp}}(\theta')]^2 d\theta'$
- The **Power Ratio** between the direct & opposite microphones:

$$\frac{\mathbb{E}_z\{[z^{\text{dir}}(t)]^2\}}{\mathbb{E}_z\{[z^{\text{opp}}(t)]^2\}} = \frac{[g^{\text{dir}}(\theta)]^2}{\bar{g}^2} \cdot \left[\frac{\sigma_d^2}{\sigma_r^2}(e^{2\delta T_r} - 1)\right] + 1$$

Directional Power Ratio (cont.)

- Replace $\mathbb{E}_z\{\cdot\} \leftrightarrow$ temporal smoothing

$\Rightarrow$ The **Directional Power Ratio quality measure**:

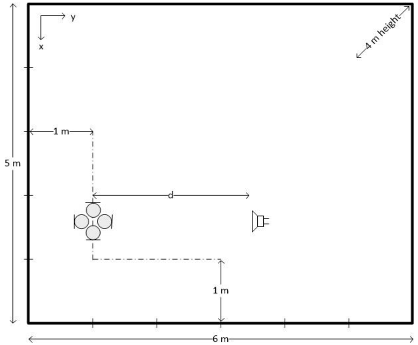
$$\mathrm{PR}(t) = \frac{P^{\mathrm{dir}}(t)}{P^{\mathrm{opp}}(t)} = \frac{\int_{t-T}^{t} [z^{\mathrm{dir}}(\tau)]^2 d\tau}{\int_{t-T}^{t} [z^{\mathrm{opp}}(\tau)]^2 d\tau} = \frac{[g^{\mathrm{dir}}(\theta)]^2}{\bar{g}^2} \cdot \mathrm{DRR}(t) + 1$$

- **Non-intrusive DRR estimator**:

$$\mathrm{PR\text{-}DRR}(t) = \frac{\bar{g}^2}{[g^{\mathrm{dir}}(\theta)]^2} \cdot \left(\frac{P^{\mathrm{dir}}(t)}{P^{\mathrm{opp}}(t)} - 1 \right)$$

Experimental Results

- Experiments:
  - **Variable source-microphone distance** with fixed T_{60}.
  - **Variable T_{60}** with fixed source-microphone distance.
- Simulation environment:



Experimental Results (cont.)

- Reference quality measures
  - Speech-to-reverberation modulation energy ratio (SRMR) (Falk, Zheng, and Chan, 2010)
  - Envelope Variance (EV) (Wolf and Nadeu, 2014)
- Correlation coefficients with:
  - *Clarity* (C50) (Kuttruff, 2009)
  - ITU-T P.862 (PESQ)
  - ITU-T P.563

Input type		White noise	Speech signals		
		Correlation ref.	Correlation ref.		
Test type	Algorithm	C50	C50	PESQ	P. 563
$T_{60} = 0.3$ sec, variable distance	PR	**0.999**	**0.999**	0.911	0.712
	SRMR	-0.27	0.845	0.973	**0.934**
	EV	-0.66	0.931	**0.994**	0.875
distance $= 0.5$ m, variable T_{60}	PR	**0.944**	**0.951**	0.899	0.562
	SRMR	0.392	0.640	**0.991**	0.873
	EV	0.235	0.614	0.984	**0.912**

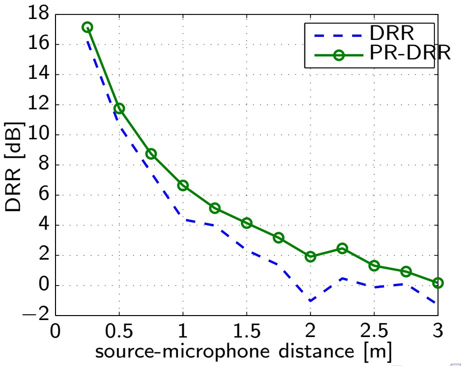
Experimental Results (cont.)

- Reference DRR measure
  - Coherent-to-diffuse-ratio (CDR)-based DRR (Jeub, Nelke, Beaugeant, and Vary, 2011)
- Correlation coefficient with:
  - DRR

Input type		White noise	Speech signals
Test type	Algorithm	Correlation ref. DRR	Correlation ref. DRR
$T_{60} = 1$ sec, variable distance	PR-DRR	**0.999**	**0.999**
	CDR	0.964	0.972
Distance = 2 m, variable T_{60}	PR-DRR	**0.999**	**0.999**
	CDR	0.852	0.913

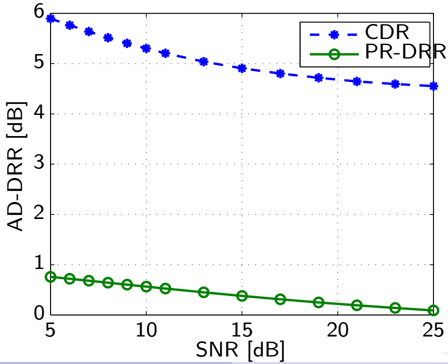
Experimental Results (cont.)

Performance of the DRR estimate for variable source-microphone distance: PR-DRR [dB] (solid-circled line), and the true DRR [dB] (dashed-line), as a function of source-microphone distance, with fixed $T_{60} = 0.3$ sec.



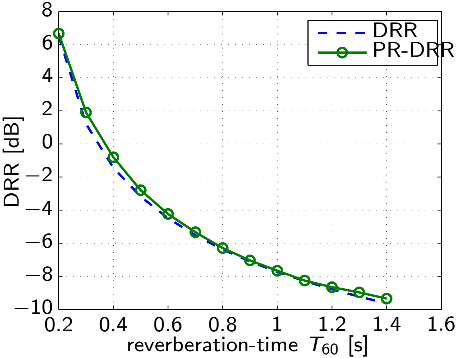
Experimental Results (cont.)

Performance of the DRR estimate for variable SNR:
Absolute difference of the proposed DRR estimate PR-DRR [dB] (solid-circled line), and of Jeub et al. CDR-based DRR estimate [dB] (dashed-asterisk line), as a function of SNR [dB]. $T_{60} = 0.3$ sec and source-microphone distance $= 0.5$ m.



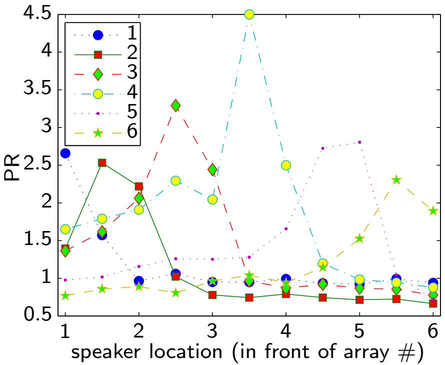
Experimental Results (cont.)

Performance of the DRR estimate for variable T_{60} - Off main-lobe:
PR-DRR [dB] (solid-circled line), and the true DRR [dB] (dashed line), as a function of T_{60}. (source–receiver angle $\in [-30^\circ .. +30^\circ]$, source-microphone distance $= 2$ m)



Experimental Results (cont.)

Recorded speech PR measure vs. source location:
The measured PR of all microphone arrays ($1-6$) vs. the source position (hall of size $15 \times 10 \times 6$ m, with 3 m spacing between adjacent arrays)



Channel Selection

- Our system is based on **clusters of uni-directional** microphones, each looking at a different direction (for demonstration, we use four uni-directional microphones looking at direction 90 degrees apart).
- We compare the signal received by each of the microphones in a cluster (referred to as local) and compare it with the other **local microphones**.

System Configuration (cont.)

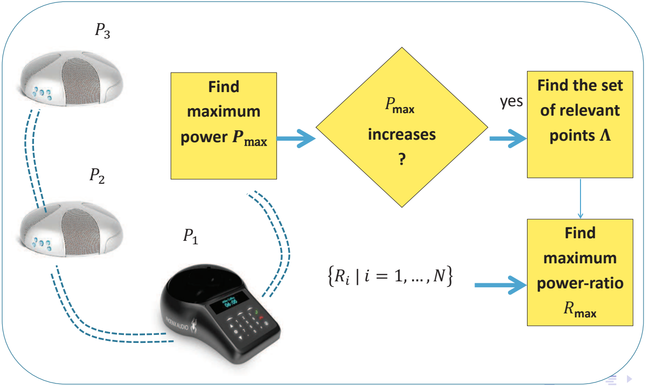
- The **PR-DRR measure** is based on the assumption that direct signals are received with different levels by the local microphones, while indirect signals (reverberations) are received with a much closer level on all the local microphones.
- We compare the PR-DRR between all the clusters and **select the audio source with the least amount of reverberation**.

Implementation

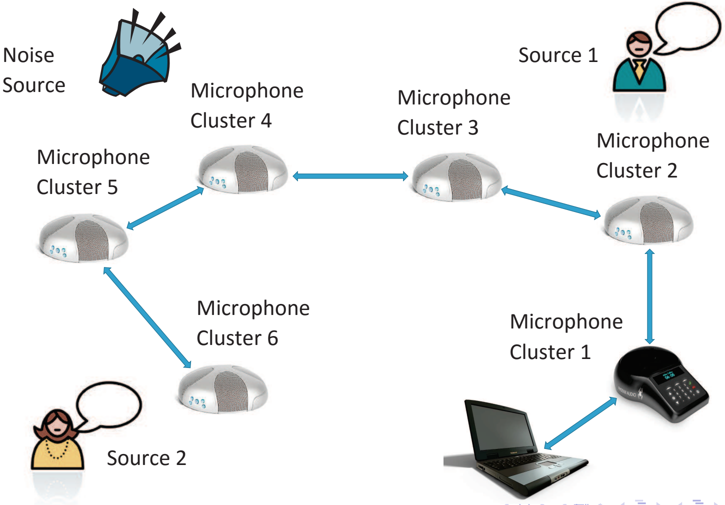
- The proposed procedure contains two stages.
  1. The first stage is **local**: for each point we compute some features of the local signals.
  2. The second stage is **global**: we select the least reverberant signal based on the features of the local signals.
- The features include **local power** and **local power-ratio**.
- The **local power** is associated with the directional microphone that measures the strongest signal at a given point, compared to the signals that are measured by the other microphones at that point.

Implementation (cont.)

- The **local power-ratio** is defined as the ratio between the local maximum power and the local minimum power.



Demonstration



Conclusions

- Instead of using randomly placed omnidirectional microphones, we use **directional microphone clusters**.
- **Calibration** is needed only within clusters, and not between clusters.
- **Short segments** of the signals are sufficient.
- The PR-DRR facilitates **fast-switching real-time selection** of the microphone with the best reception amongst randomly placed microphone clusters in a conference room.

Future Work

- Directional non-stationary noise.
- Time delay between signals in different clusters.
- Direction of arrival estimation.
- Clusters of circular differential microphone arrays.
- Combine the PR-DRR with other measures (e.g., spatial coherence).