Speech Dereverberation based on Convex Optimization Algorithms for Group Sparse Linear Prediction

AALBORG UNIVERSITY
DENMARK

SONOS
The Home Sound System

Daniele Giacobello[1], Tobias Lindstrøm Jensen[2]

[1]Sonos Inc., Santa Barbara, CA, USA

[2]Signal and Information Processing, Dept. of Electronic Systems, Aalborg University, Denmark

Motivation

- Speech dereverberation fundamental for enabling far-field human-computer interaction, particularly with the recent advent of *smart* loudspeaker devices (e.g., Sonos One ☺).
- Blind methods based on multi-channel linear prediction (MCLP) applied in the STFT-domain particularly effective for the task:
  - no prior knowledge of the room acoustics,
  - relatively easy and cheap to implement.
- Popular MCLP-based methods look for a *sparse* desired speech signal, assuming reverberation as a convolutive process (approximated by the predicted speech) on a STFT bin-by-bin basis. This is done by applying nonconvex algorithms [1, 2].
- We propose alternative formulations for sparse approximation based on convex optimization [3].

MCLP-based Dereverberation

- Focus on *utterance-based batch processing*
- Reverberant speech signal model at $m-$th mic $m \in \{1, \dots, M\}$:

$$x_m(k,n) = \underbrace{\sum_{l=0}^{\tau-1} h_m(k,l)s(k,n-l)}_{d_m(k,n)} + \underbrace{\sum_{l=\tau}^{L_g-1} h_m(k,l)s(k,n-l)}_{r_m(k,n)} \quad (1)$$

  - $n \in \{1, \dots, N\}$ frame index, $k \in \{1, \dots, K\}$ frequency bin index
  - $s(k,n)$: clean speech
  - $d_m(k,n)$: desired speech
  - $r_m(k,n)$: reverberation term
  - $h_m(k,l)$ ATF between the speech source and the m-th microphone
  - τ: delay to model direct speech and early reflections
  - L_g: prediction order
- Desired speech signal using M predictors (order (L_g-1)):

$$d_m(k,n) = x_m(k,n) - \sum_{i=1}^{M}\sum_{l=0}^{L_g-1} x_i(k,n-\tau-l)g_{m,i}(k,l) \quad (2)$$

  - $g_{m,i}(k,l)$: $l-$th prediction coefficient between the $i-$th and the $m-$th channel

Group Sparse Linear Prediction

- The model in (2) in matrix form becomes [2]:

$$\mathbf{D}(k) = \mathbf{X}(k) - \mathbf{X}_\tau(k)\mathbf{G}(k) \quad (3)$$

$$\begin{aligned}
\text{with}\quad \mathbf{D}(k) &= [\mathbf{d}_1(k), \cdots, \mathbf{d}_M(k)] \in \mathbb{C}^{N\times M}\\
\mathbf{d}_m(k) &= [d_m(k,1), \cdots, d_m(k,N)]^T \in \mathbb{C}^{N\times 1}\\
\mathbf{X}(k) &= [\mathbf{x}_1(k), \cdots, \mathbf{x}_M(k)] \in \mathbb{C}^{N\times M}\\
\mathbf{x}_m(k) &= [x_m(k,1), \cdots, x_m(k,N)]^T \in \mathbb{C}^{N\times 1}\\
\mathbf{X}_\tau(k) &= [\mathbf{X}_{\tau,1}(k), \cdots, \mathbf{X}_{\tau,M}(k)] \in \mathbb{C}^{N\times ML_g}\\
\mathbf{G}(k) &= [\mathbf{g}_1(k), \cdots, \mathbf{g}_M(k)] \in \mathbb{C}^{ML_g\times M}\\
\mathbf{g}_m(k) &= [g_{m,1}(k,0), \cdots, g_{m,1}(k,L_g-1), \dots\\
&\quad g_{m,M}(k,0), \cdots, g_{m,M}(k,L_g-1)]^T \in \mathbb{C}^{ML_g\times 1}
\end{aligned}$$

- $\mathbf{G}$ in (3) is then found by solving the optimization problem:

$$\hat{\mathbf{G}} = \underset{\mathbf{G}}{\operatorname{argmin}} \| \mathbf{X} - \mathbf{X}_\tau\mathbf{G} \|_{p,1}^1 + \alpha \|\mathbf{G}\|_{1,1}^1 \quad (4)$$

  - $\|\cdot\|_{p,1}^1$: $\mathbf{V} \in \mathbb{C}^{n\times m}$, $\|\mathbf{V}\|_{p,1} = \left(\sum_{i=1}^n \|\mathbf{V}_{i,:}\|_p\right)$
  - $\|\mathbf{V}_{i,:}\|_p$ is the ℓ_p norm of the $i-$th row-vector $\mathbf{V}_{i,:}$
  - For $p=1$, (4) is a element-wise regularized least-sum-of-absolute
  - For $p=2$, (4) is a group LASSO problem
  - $\alpha\|\mathbf{G}\|_{1,1}^1$ regularization term meaning:
    - ⋆ ill-conditioning when closed-spaced microphones: $\mathbf{X}_\tau^H\mathbf{X}_\tau \to$ singular
    - ⋆ model order selection penalization if L_g is not chosen appropriately

Experimental Setup

- 6-microphone circular array of 72 mm diameter
- Performance evaluated by simulating artificial utterances that mimic real use cases specific for voice enabled smart speakers:
  - Room size: $w \in [3,8]$ m, $l \in [3,10]$ m $h \in [2,4]$ m
  - Position: $d \in [1,7]$ m, azimuth $\theta \in [-180, 180]$, elevation $\phi \in [45, 135]$
  - Tuned reflection coefficients of cuboid to obtain $T_{60} \in [300, 700]$ ms
  - COMSOL® used to solving the scalar wave equation using the finite element method
- diffuse HVAC noise, SNR $\in [10, 30]$ dB (focus on dereverberation)
- ASR engine trained using the *Librispeech 100hrs* corpus: 100 hours of *clean* speech, 125 male, 125 female speakers), audiobooks data
- STFT 50% overlap 32ms Hamming (f_s = 16kHz). $L_g = 10$, $\tau = 2$
- 100 iterations of ADMM, 5-7 iterations IRLS

Convex Formulations

- Non-convex formulation (Iteratively Reweighted Least-Squares)
  - ℓ_q norm $(0 < q \leq 1)$ approximated using IRLS. $i-$th step:

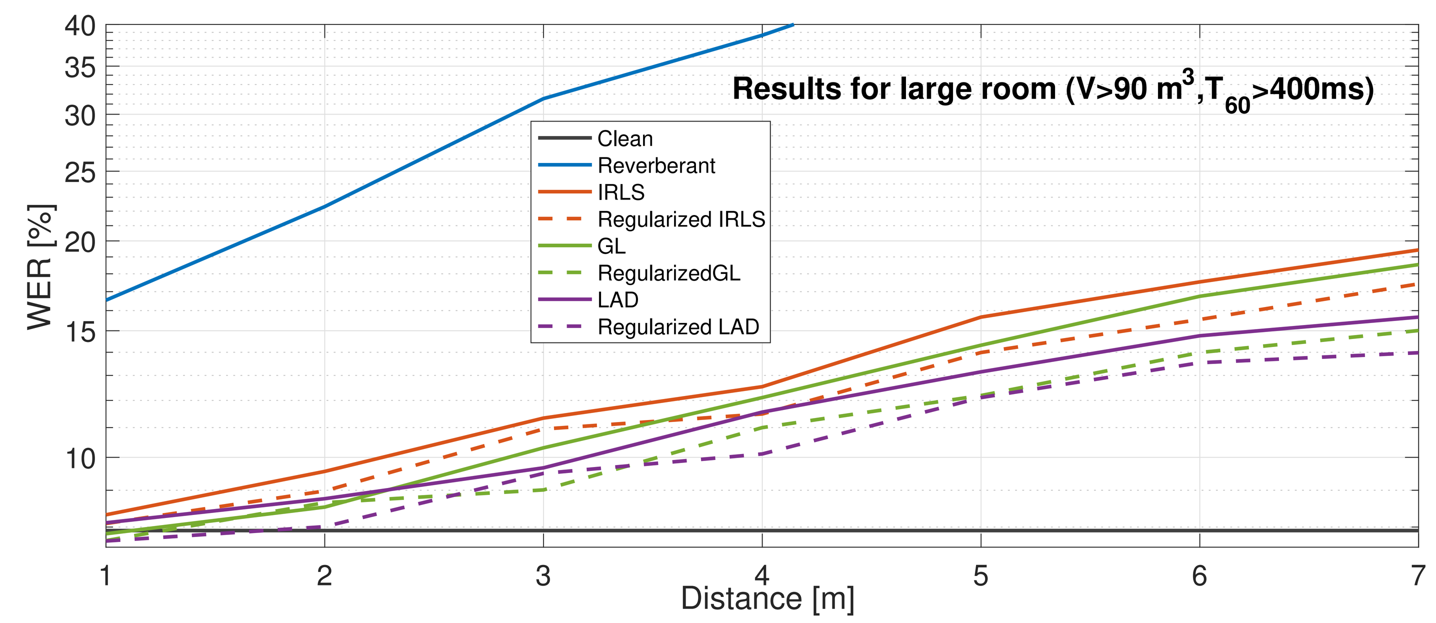
$$\hat{\mathbf{G}}^i = \underset{\mathbf{G}}{\operatorname{argmin}} \|\mathbf{W}_i^{1/2}(\mathbf{X} - \mathbf{X}_\tau\mathbf{G})\|_2^2$$

  - with $\mathbf{W}^i = \operatorname{diag}(\mathbf{w}^i)$, $\mathbf{w}^i = \left(\|\mathbf{d}_n\|_2^2 + \epsilon\right)^{q/2-1}$, $\forall n$, updated from $\hat{\mathbf{D}}^i$
- Least Absolute Devation (LAD)
  - $p = 1$ in (4), problem is separable:

$$\hat{\mathbf{g}}_m = \underset{\mathbf{g}_m}{\operatorname{argmin}} \|\mathbf{x}_m - \mathbf{X}_\tau\mathbf{g}_m\|_1 + \alpha\|\mathbf{g}_m\|_1, \quad m = 1, \dots, M$$

  - known ADMM formulation [3] for $\hat{\mathbf{g}}_m = \operatorname{argmin}_{\mathbf{g}_m} \left\| \begin{bmatrix} \mathbf{x}_m \\ 0 \end{bmatrix} - \begin{bmatrix} \mathbf{X}_\tau \\ \alpha\mathbf{I} \end{bmatrix} \mathbf{g}_m \right\|_1$
- Group LASSO (GL)
  - $p = 2$ in (4), problem is non separable, $i-$th ADMM steps are:
    1. $\hat{\mathbf{G}}^i = (\mathbf{X}_\tau^H\mathbf{X}_\tau + \alpha\mathbf{I})^{-1}\left[\alpha\mathbf{I} \quad \mathbf{X}_\tau^H\right]\left(\mathbf{Z}^i + \begin{bmatrix}\mathbf{0}\\ \mathbf{X}\end{bmatrix} - \mathbf{L}^i\right)$
    2. $\mathbf{R}^i = \begin{bmatrix}\alpha\hat{\mathbf{G}}^i \\ \mathbf{X}_\tau\hat{\mathbf{G}}^i - \mathbf{X}\end{bmatrix}$
    3. $\mathbf{Z}^{i+1} = \mathcal{S}_t(\mathbf{R}^i + \mathbf{L}^i)$
    4. $\mathbf{L}^{i+1} = \mathbf{L}^i + \mathbf{R}^i - \mathbf{Z}^{i+1}$
  - where the proximity operator $\mathcal{S}_t(\cdot)$ subproblem is separable
- Complexity: IRLS $\mathcal{O}((ML_g)^3 + N(ML_g)^2)$, ADMM $\mathcal{O}(M^3L_g^2)$
- Code available at: `https://github.com/giacobello/`

Results



References

[1] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, 2010.

[2] A. Jukić, T. van Waterschoot, T. Gerkmann, and S. Doclo, "Group sparsity for MIMO speech dereverberation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2015.

[3] T. L. Jensen, D. Giacobello, T. van Waterschoot, and M. G. Christensen, "Fast algorithms for high-order sparse linear prediction with applications to speech processing," *Speech Communication*, vol. 76, pp. 143 – 156, 2016.

Acknowledgments

This work (including D. Giacobello research stay at Aalborg University) was partly supported by the Danish Council for Independent Research, grant no. 4005-00122.