A Computationally Constrained Optimization Framework for Implementation and Tuning of Speech Enhancement Systems

Daniele Giacobello, Jason Wung, Ramin Pichevar, Joshua Atkins

Beats Electronics, Culver City, CA

Contact Information:
Beats Electronics
8600 Hayden Place
Culver City, CA 90232

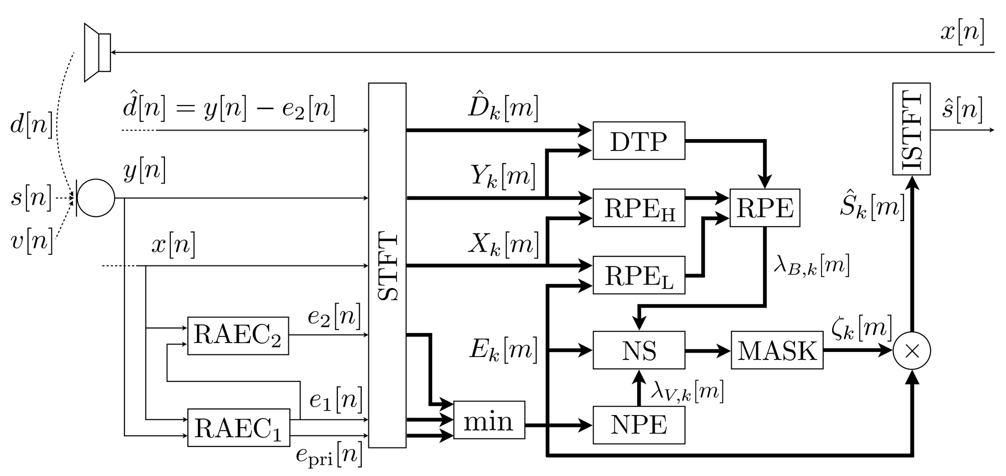
Email: {dgiacobello,jwung}@apple.com

Motivation

- Speech enhancement (SE) systems integrate different algorithms and aim at maximizing their overall performance using objective measures:
  - Mean Opinion Score (MOS) for full-duplex communication schemes.
  - Phone Accuracy Ratio (PAR) for ASR front-ends.
- Commercially viable SE system must take into account the computational budget of the target hardware.
- Procedure for tuning the parameters of an SE system $\mathbf{p} = \{p_1, p_2, \ldots, p_N\}$ are not explicitly formalized and highly suboptimal:
  - Each component profiled separately.
  - Use of measures easier to handle but not related to the actual overall target (e.g., MSE).
  - Tuning only done at an advanced stage of the development relying on small test cases.

1 Speech Enhancement System

1.1 Architecture



Block diagram of the speech enhancement system.

- **Robust Acoustic Echo Canceler** (RAEC) employs an error recovery nonlinearity allowing for continuous update. Multi-delay adaptive filter structure (1,2).
- **Residual Echo Power Estimator** (RPE) based on coherence (3,4).
- **Double Talk Probability** (DTP) based on coherence (5).
- **Noise Power Estimator** (NPE) based on (6), implicitly accounting for the speech presence probability (SPP).
- **Direct Masking** (MASK) applies a masking based on (8) or *quasi*-binary based on (9) depending on the SNR.

1.2 Complexity Analysis

- While the actual complexity is platform dependent, each fundamental operations can be estimated in terms of DSP cycles, thus subsequently calculated in terms of million cycles per second (MCPS).
- Dividing the analysis per sample for each block

$$
\begin{aligned}
C_{\text{RAEC}} &= (3N_{\text{iter}}+2)\text{-FFT}_{\text{RAEC}} + (5N_{\text{iter}}+3)\text{-mply} + (3N_{\text{iter}}+1)\text{-MAC} \\
&\quad + (2N_{\text{iter}}+1)\text{-cplx-pwrSpectr} + (2N_{\text{iter}}+1)M_{\text{RAEC}}\text{-cplx-mply} \\
&\quad + N_{\text{iter}}(M_{\text{RAEC}}+1)\text{-add} + N_{\text{iter}}\text{-sqrt} + 2N_{\text{iter}}\text{-div} + N_{\text{iter}}\text{-if-else} \\
&\quad + N_{\text{iter}}M_{\text{RAEC}}\text{-real-cplx-mply} \\
C_{\text{STFT}} &= 2\text{-mply} + \text{FFT}_{\text{STFT}} \\
C_{\text{DTP}} &= 3\text{-cplx-pwrSpectr} + 18\text{-mply} + 12\text{-MAC} + 1\text{-cplx-mply} + 6\text{-div} \\
&\quad + 9\text{-add} + 1\text{-exp} + 1\text{-sqrt} + 1\text{-log} \\
C_{\text{RPE}} &= 1\text{-cplx-pwrSpectr} + 4\text{-mply} + 3\text{-MAC} + (M_{\text{RPE}}+1)\text{-cplx-mply} \\
&\quad + (M_{\text{RPE}}+1)\text{-add} + 1\text{-div} \\
C_{\text{NPE}} &= 1\text{-cplx-pwrSpectr} + 3\text{-div} + 3\text{-add} + 5\text{-mply} + 1\text{-exp} + 3\text{-MAC} \\
&\quad + 2\text{-if-else} \\
C_{\text{NS}} &= 2\text{-cplx-pwrSpectr} + 2\text{-add} + 1\text{-if-else} + 3\text{-mply} + 2\text{-MAC} + 3\text{-div}
\end{aligned}
$$

- The overall complexity of the system is then

$$
\begin{aligned}
\mathrm{C}(\mathbf{p}) = (&C_{\text{RAEC}_1} + C_{\text{RAEC}_2} + 7C_{\text{STFT}} + C_{\text{DTP}} \\
&+ C_{\text{RPE}_\text{H}} + C_{\text{RPE}_\text{L}} + C_{\text{NPE}} + C_{\text{NS}})\frac{f_s}{10^6} \text{ (MCPS)}.
\end{aligned}
$$

- Note:
  - The tuning parameters highlighted above are the one affecting directly the computational cost.
  - Defined binary parameters that enable/disable algorithmic components.
  - Other parameters, e.g., smoothing factors, time constants, and thresholds, should also be optimized jointly.

2 Optimization Framework

- The tuning problem can be formulated mathematically as a constrained optimization problem.
- Let $\hat{s}[n, \mathbf{p}]$ be the SE system output obtained with $\mathbf{p}$, the problem can be written as:

$$
\boxed{\begin{aligned} &\text{maximize } \mathrm{Q}(\hat{s}[n,\mathbf{p}]), \\ &\text{subject to } \mathrm{C}(\mathbf{p}) \le C_{\text{max}}. \end{aligned}}
$$

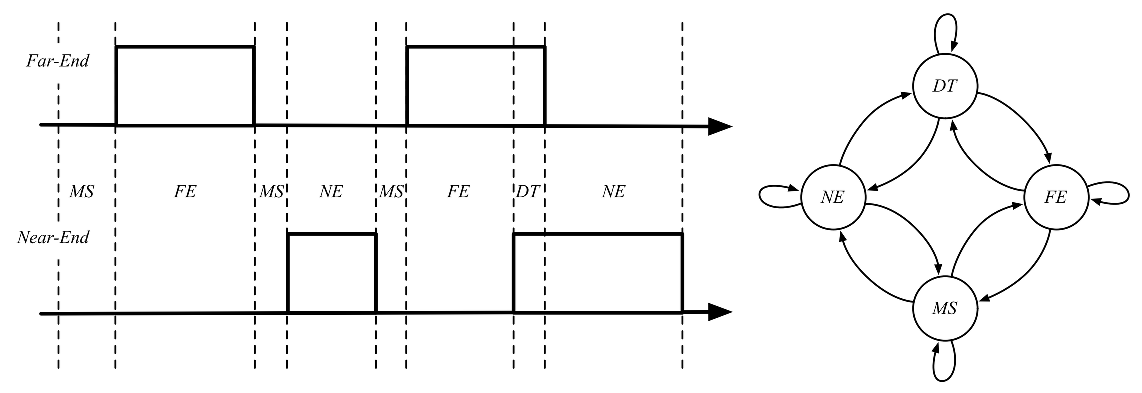
  where $\mathrm{Q}(\cdot)$ is the optimization criterion and C_{max} is the computational complexity constraint.
- We choose to solve this nonlinear programming problem applying a genetic algorithm (GA). Using operators such as *mutation* and *crossover* are used to evolve a set of solutions, $\mathbf{\Pi}^{(k)} = \{\mathbf{p}_m^{(k)}, m = 1, \ldots, M\}$. At convergence ($K$ iterations), we obtain:

$$
\hat{\mathbf{p}} = \arg\max_{\mathbf{p}_m^{(K)} \in \mathbf{\Pi}^{(K)}} \mathrm{Q}\left(\hat{s}[n, \mathbf{p}_m^{(K)}]\right) \quad \text{s.t.} \quad \mathrm{C}(\mathbf{p}_m^{(K)}) \le C_{\text{max}}.
$$

3 Experimental Analysis

3.1 Dataset Generation

- Key element for the proposed approach is to have a well structured database for training and testing that correlates well with real world scenarios.
- We applied statistics of conversational speech to generate a database of 3,150 conversational sequences from the TIMIT database for training and 3,150 for testing (length between 6 to 8 s).
- Signal-to-Echo Ratio (SER) was uniformly distributed between -30 and 5 dB and 10 RIRs were used, measured in office environments.
- Signal-to-Noise Ratio (SNR) uniformly distributed between -5 to 10 dB (different types of noise).



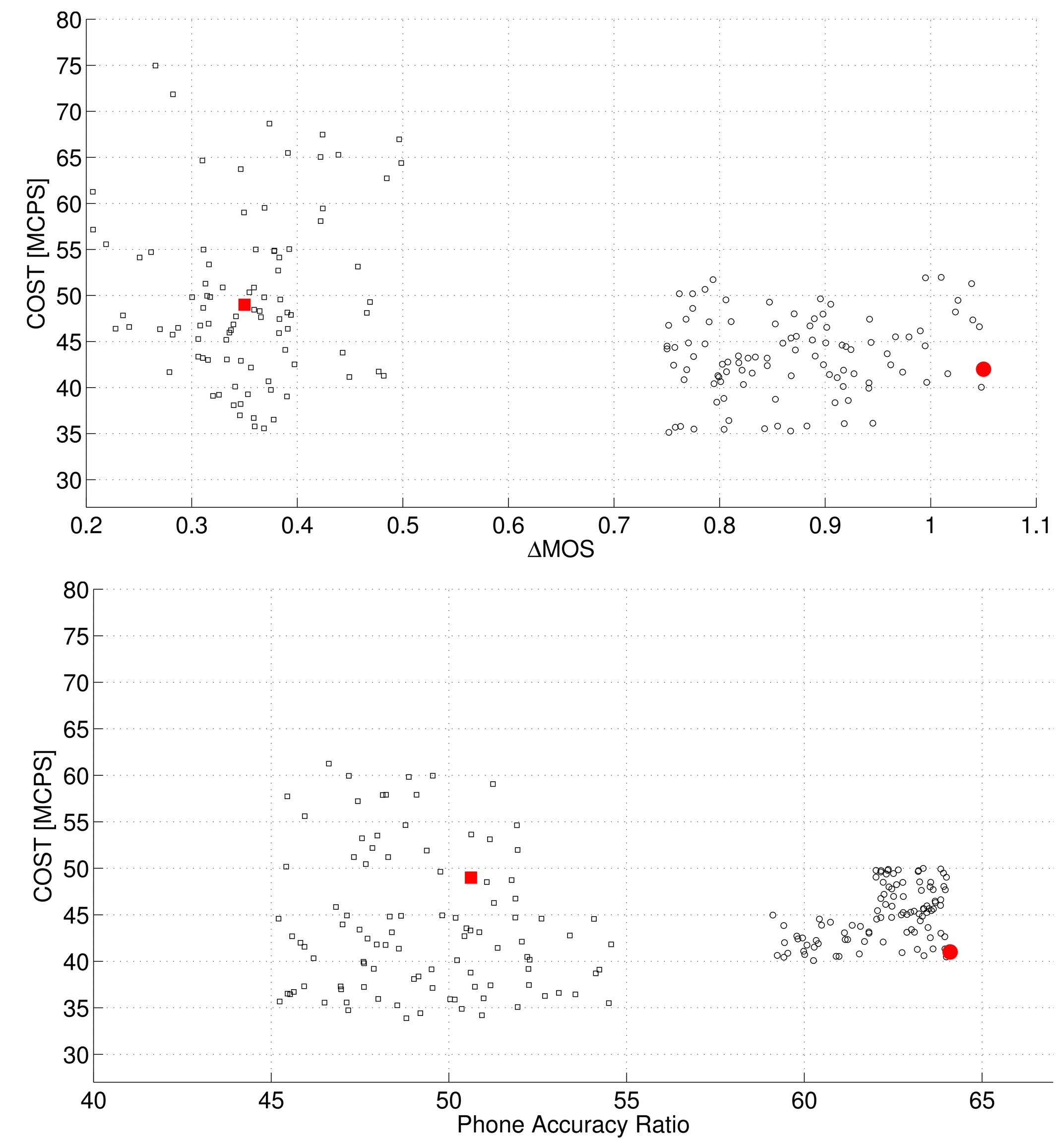
Example of a conversational speech sequence and its Markov chain generative model.

3.2 Setup and Results

- Optimization criteria:
  - Median $\Delta\text{MOS}\,(\hat{s}[n,\mathbf{p}], y[n])$ obtained through POLQA (10), calculated for each utterance and averaged over training set.
  - PAR calculated over training set using acoustic model of 61 phones, 13 MFCCs +13 ΔMFCCs + 13 $\Delta\Delta$MFCCs, 5-state HMMs, 8-mixture GMMs (training on clean speech only to focus on SE (9)).
- Constraint: $C_{\text{max}} = 50$ MCPS
- GA with population of 100 elements and 10 generations run (convergence reached); 90 hours on a 16-core Intel Xeon machine with parallelized scripts.

Results of the GA optimization algorithm (test TIMIT) (constrained vs. unconstrained).

	PAR (%)	ΔMOS	$\mathrm{C}(\mathbf{p})$ (MCPS)
$\mathbf{p}_{\text{INIT}}$	51.04	0.32	49.14
$\hat{\mathbf{p}}_{\text{PAR}}$	62.94	0.65	41.17
$\hat{\mathbf{p}}_{\text{PAR}_\text{u}}$	63.15	0.68	53.56
$\hat{\mathbf{p}}_{\text{MOS}}$	60.07	0.87	42.56
$\hat{\mathbf{p}}_{\text{MOS}_\text{u}}$	60.22	0.92	55.23



Initial population (squares) and final population (circles) of the GA in the constrained optimization over ΔMOS and PAR on the training database. The initial solution $\mathbf{p}_{\text{INIT}}$ is the red square, while the optimal final solution that respects the constraint is the red circle.

4 Conclusions

- Results over presented SE system showed:
  - Net improvement over an initial solution hand-tuned by an expert both in terms of MOS (+0.55) and PAR (+11.90%).
  - Complexity kept below imposed target of 50 MCPS (20% less complex than initial solution).
- Proposed system can be very helpful in the prototyping phase as well as in the conceptual stage of algorithmic design.

References

(1) J. Wung, T. S. Wada, B.-H. Juang, B. Lee, M. Trott, and R. W. Schafer, "A System Approach to Acoustic Echo Cancellation in Robust Hands-Free Teleconferencing," *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 101–104, 2011.

(2) T. S. Wada and B.-H. Juang, "Enhancement of Residual Echo for Robust Acoustic Echo Cancellation," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 175–189, 2012.

(3) G. Enzner, R. Martin, and P. Vary, "Unbiased Residual Echo Power Estimation for Hands-Free Telephony," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, pp. 1893–1896, 2002.

(4) S. Goetze, M. Kallinger, and K.–D. Kammeyer, "Residual Echo Power Spectral Density Estimation Based on an Optimal Smoothed Misalignment For Acoustic Echo Cancellation," *Proc. International Workshop on Acoustic Echo and Noise Control*, pp. 209–212, 2005.

(5) I. J. Tashev, "Coherence Based Double Talk Detector with Soft Decision," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 165–168, 2012.

(6) T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-Based Noise Power Estimation with Low Complexity and Low Tracking Delay," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393, 2012.

(7) Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

(8) ——, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

(9) R. Pichevar, A. Ziaei, J. Wung, D. Giacobello, and J. Atkins, "Design and Optimization of a Speech Recognition Front-End for Distant-Talking Control of a Music Playback Device," submitted to *5th IEEE Workshop on Spoken Language Technology*, 2014.

(10) *Perceptual Objective Listening Quality Assessment*, ITU-T Rec. P.863, 2010.