Tuning Methodology for Speech Enhancement Algorithms using a Simulated Conversational Database and Perceptual Objective Measures

Daniele Giacobello, Jason Wung, Ramin Pichevar, and Joshua Atkins

Beats Electronics, LLC, Santa Monica, CA

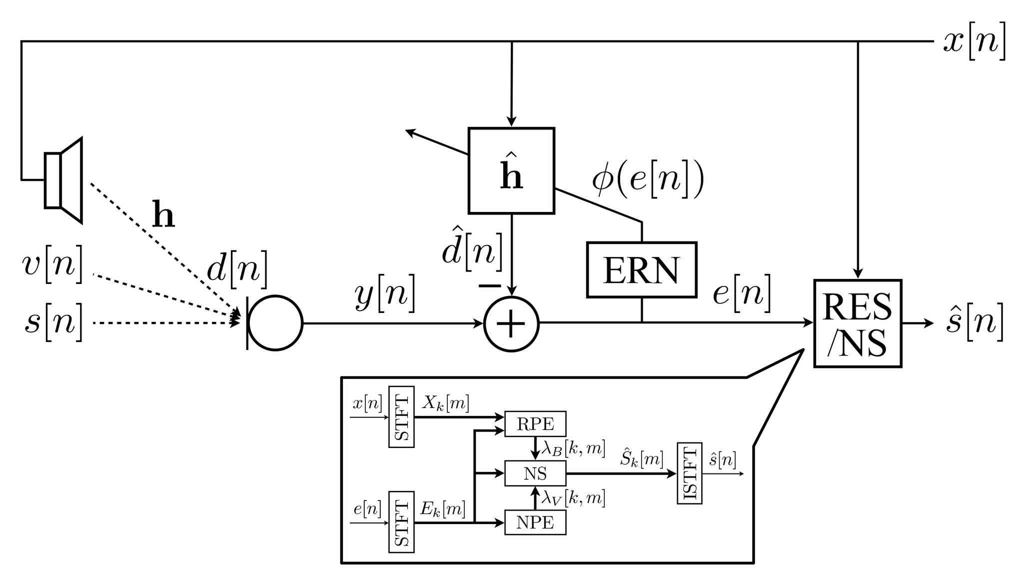
Contact Information:
Beats Electronics, LLC
1601 Cloverfield Blvd.
Suite 5000N
Santa Monica, CA 90404, USA

Email: daniele.giacobello@beatsbydre.com

Motivation

- Tuning is critical for real-world deployment of speech enhancement (SE) systems for full-duplex communications:
  - Impractical designing algorithms that try to cover all possible interferences.
  - System is often hand-tuned by experts and needs to be verified through subjective listening tests.
  - Hand-tuning process is time-consuming, error-prone, and bound to cover only a relatively small number of scenarios.
- Tuning procedure is not explicitly formalized (1):
  - Combinatorial nature of the problem.
  - Optimization criteria relates to the fuzzy concept of *perceptual quality*.
  - Need for a realistic training and testing large database.

1 Speech Enhancement System



Block diagram of the speech enhancement system.

- **Robust Acoustic Echo Canceler** (RAEC) (2,3) with an error recovery nonlinearity allowing for continuous update. Multi-delay adaptive filter structure.
  - Tuning parameters: number of partitioned blocks M_{AEC}, number of iterations N_{AEC}, the step-size μ_{AEC}, and the smoothing factor α_{AEC} for the power spectral density estimation.
- **Residual Echo Power Estimator** (RPE) based on coherence (4,5).
  - Tuning parameters: number of past frames M_{RPE} and the smoothing factor α_{RPE}.
- **Noise Power Estimator** (NPE) based on (6), implicitly accounting for the speech presence probability (SPP).
  - Tuning parameters: the fixed *a priori* SNR ξ_{H_1}, the SPP threshold P_{TH}, and the smoothing factors α_P and α_{NPE}.
- **Noise Suppressor** (NS) based on log-spectral amplitude MMSE estimator (7) with estimation of the *a priori* SNR using decision-directed approach (8).
  - Tuning parameters: smoothing factor for the SNR estimator α_{DD} and the minimum suppression gain G_{min}.

2 Tuning as an Optimization Problem

- The objective of a SE algorithm is to *maximize* the quality of the speech output $\hat{s}[n,\mathbf{p}]$, obtained with the set of tunable parameters $\mathbf{p}$.
- Since measures are full-referenced, we calculate the *difference* in MOS as

$$\Delta\text{MOS}\left(\hat{s}[n], y[n]\right) = \text{MOS}\left(\hat{s}[n], s[n]\right) - \text{MOS}\left(y[n], s[n]\right).$$

- Imposing simple bounds on the parameter values, the problem becomes:

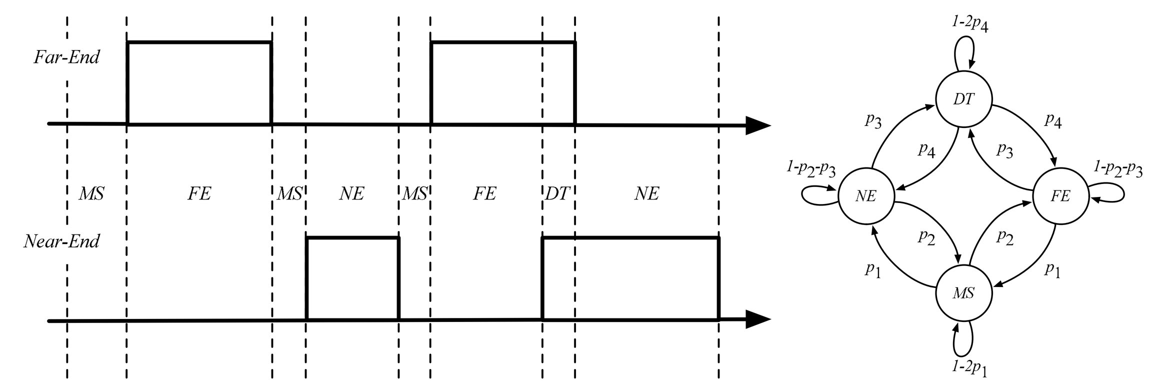
$$\begin{array}{|l|}\hline \text{maximize} \quad \Delta\text{MOS}\left(\hat{s}[n,\mathbf{p}], y[n]\right) \\ \text{subject to } \mathbf{U} \le \mathbf{p} \le \mathbf{L}. \\ \hline\end{array}$$

- We choose to solve this nonlinear programming problem applying a genetic algorithm. 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:

$$\hat{\mathbf{p}} = \underset{\mathbf{p}_m^{(K)} \in \mathbf{\Pi}^{(K)}}{\arg\max}\ \Delta\text{MOS}\left(\hat{s}[n, \mathbf{p}_m^{(K)}], y[n]\right).$$

3 Database Generation

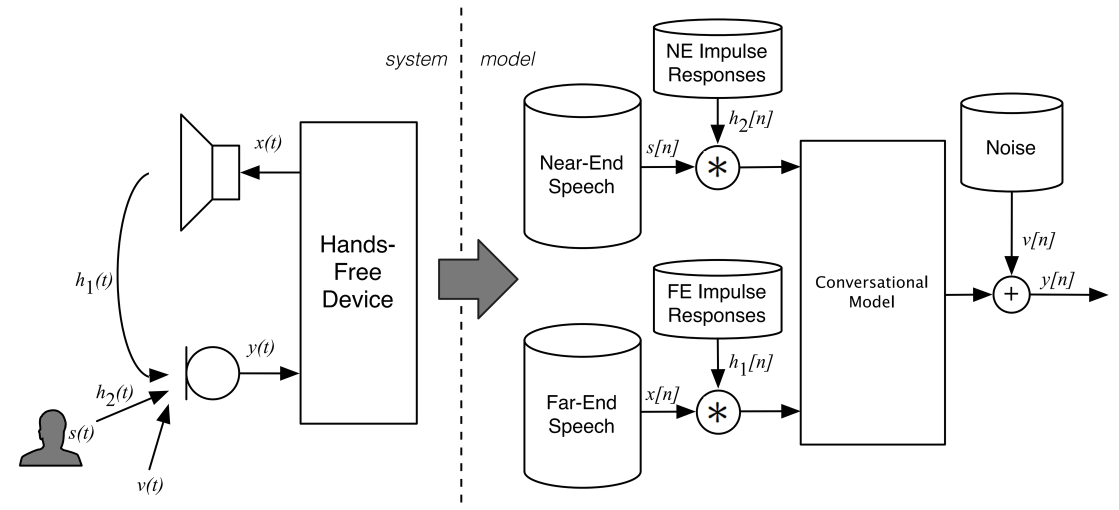
- A proper database is necessary to determine reliable solutions.
- Modeling of human conversational speech and conversational events, as proposed in (9) is rather simplistic and relies on hand-coded expert knowledge.
- We use a 4-state Markov chain based on the probabilities defined in (9) to find a flexible solution for automatic generation of a large conversational speech database.
- Can be easily modified to fit different types of conversation scenarios with different levels of interactivity.



Conversational sequence and its Markov chain model.

4 Experimental Analysis

4.1 Setup



Model of the loudspeaker-microphone configuration.

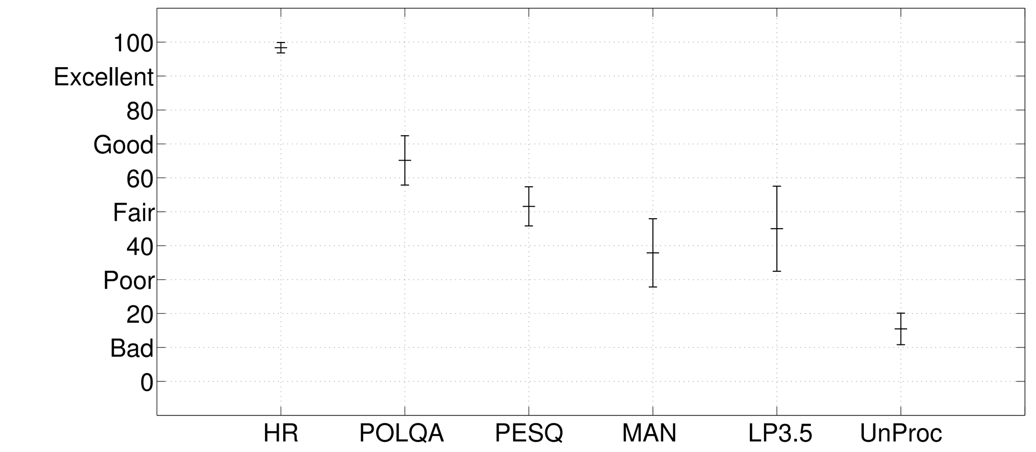
- Two single-channel signals, NE and FE, with continuous activity (i.e., without pauses) were generated from the ITU-T P-Series test signals.
- We generated 1000 segments with lengths between 6 to 8 s, ideal for objective quality measures, choosing randomly starting and ending point in the FE and NE signals.
- 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).
- 80% of database used for training, 20% for testing.

4.2 Results

- ΔMOS was obtained through PESQ (10), POLQA (11), and ViSQOL (12).
- The optimization framework was also used with objective measures, averaged over the evaluation set, that do not account for perception: LSD, tERLE, MSE, and tERLE (for AEC) + LSD (for RPE, NPE, and NS).

Comparison between the objective improvements obtain with the SE algorithm in terms of MOS calculated with POLQA, PESQ, and ViSQOL obtained with different sets of parameters as result of optimizing with different criteria. A 95% confidence interval is given for each value.

method	$\Delta\text{MOS}_{\text{PESQ}}$	$\Delta\text{MOS}_{\text{POLQA}}$	$\Delta\text{MOS}_{\text{ViSQOL}}$
$\mathbf{p}_{\text{POLQA}}$	.455±.021	.654±.042	.387±.021
$\mathbf{p}_{\text{PESQ}}$	.475±.035	.442±.050	.342±.053
$\mathbf{p}_{\text{ViSQOL}}$	.358±.028	.487±.450	.369±.032
$\mathbf{p}_{\text{MANUAL}}$	.276±.083	.296±.121	.201±.089
$\mathbf{p}_{\text{LSD}}$	.139±.042	.221±.046	.154±.043
$\mathbf{p}_{\text{tERLE}}$	.147±.053	.234±.067	.121±.025
$\mathbf{p}_{\text{tERLE+LSD}}$	.194±.061	.246±.049	.173±.082
$\mathbf{p}_{\text{MSE}}$	.138±.089	.179±.134	.104±.091



MUSHRA listening test results comparing six speech samples obtained with different optimization criteria. A pool of eleven expert listeners, familiar in detecting small impairments, and seven naive listeners was chosen.

5 Conclusions

- The use of perceptual objective measures for large-scale optimization greatly improves the performance of the SE algorithm over a much larger dataset than commonly used.
- $\Delta\text{MOS}_{\text{POLQA}}$ shows that $\mathbf{p}_{\text{POLQA}}$ is .358 above $\mathbf{p}_{\text{MANUAL}}$ which is remarkable since there is no algorithmic modification other than using a better perceptual objective measure.

References

(1) I. Tashev, A. Lovitt, and A. Acero, "Unified Framework for Single Channel Speech Enhancement," *Proc. IEEE Pacific Rim Conference on Communications, Computers, and Signal Processing*, pp. 883–888, 2009.

(2) 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.

(3) 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.

(4) G. Enzner, R. Martin, and P. Vary, "Unbiased residual echo power estimation for hands-free telephony," *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 2, pp. 1893–1896, May 2002.

(5) 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.

(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) *Artificial Conversational Speech*, ITU-T Rec. P. 59, 1993.

(10) *Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs*, ITU-T Rec. P.862, 2001.

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

(12) A. Hines, J. Skoglund, A. Kokaram, and N. Harte, "ViSQOL: The Virtual Speech Quality Objective Listener," *Proceedings of the International Workshop on Acoustic Echo and Noise Control*, pp. 1–4, 2012.