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

Motivation

- Tuning is critical for real-world deployment of speech enhancement (SE) systems for full-duplex communications.
- Implicitly 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): Combinational 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

- Robust Acoustic Echo C canceler (RAEC) (2, 3) with an error recovery nonlinearity allowing for continuous update. Multi-delay adaptive filter structure.
- Tuning parameters: number of partitioned blocks $N_{\text{RAEC}}$, number of iterations $N_{\text{RAEC}}$, the step-size $\eta_{\text{RAEC}}$, and the smoothing factor $\alpha_{\text{RAEC}}$ for the power spectral density estimation.
- Residual Echo Power Estimator (RPE) based on coherence (4, 5).
- Tuning parameters: number of past frames $N_{\text{RPE}}$ and the smoothing factor $\alpha_{\text{RPE}}$.
- Noise Power Estimator (NPE) based on (6), implicitly accounting for the speech presence probability (SPP).

2 Tuning as an Optimization Problem

- The objective of a SE algorithm is to maximize the quality of the speech output $s[n, p]$, obtained with the set of tunable parameters $p$.
- Since measures are full-referenced, we calculate the difference in MOS as $\Delta \text{MOS}(s[n], y[n]) = \text{MOS}(s[n], s[n]) - \text{MOS}(y[n], s[n])$.
- Imposing simple bounds on the parameter values, the problem becomes:

$$\begin{align*}
\text{maximize} & \quad \Delta \text{MOS}(s[n], p; y[n]) \\
\text{subject to} & \quad U \leq p \leq L
\end{align*}$$

- 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, $P^{(k)} = (p^{(k)}_{m}, m = 1, \ldots, M)$. At convergence ($K$ iterations), we obtain:

$$p = \arg \max_p \Delta \text{MOS}(s[n, p], y[n])$$

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 conversational scenarios with different levels of interactivity.

4 Experimental Analysis

4.1 Setup

- 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 IRs 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

- $\Delta \text{MOS}$ 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, IERLE, MSE, and IERLE (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.

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 $\text{POLQA}$ is .358 above $\text{MANUAL}$ which is remarkable since there is no algorithmic modification other than using a better perceptual objective measure.

References