# TUNING METHODOLOGY FOR SPEECH ENHANCEMENT ALGORITHMS USING A SIMULATED CONVERSATIONAL DATABASE AND PERCEPTUAL OBJECTIVE MEASURES

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

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

{Daniele.Giacobello, Jason.Wung, Ramin.Pichevar, Josh.Atkins}@beatsbydre.com

## ABSTRACT

In this paper, we propose a formal methodology for tuning the parameters of a single-microphone speech enhancement system for hands-free devices. The tuning problem is formulated as a large-scale nonlinear programming problem that is solved by a genetic algorithm to determine the global solution. A conversational speech database is automatically generated by modeling the interactivity in telephone conversations, and perceptual objective quality measures are used as the optimization criteria for the automated tuning over the generated database. A subjective listening test is then performed by comparing the automatically tuned system based on objective criteria to the system tuned by expert human listeners. Subjective and objective evaluation result shows that the proposed automated tuning methodology greatly improves the enhanced speech quality, potentially saving resources over manual evaluation, speeding up development and deployment time, and guiding the algorithmic design.

***Index Terms—*** Acoustic Echo Cancellation, Speech Enhancement, Conversation Analysis, Perceptual Objective Quality.

## 1. INTRODUCTION

Speech enhancement (SE) algorithms are fundamental to a large number of speech-centric applications, such as mobile communication, speech recognition, and hearing aids [1], especially when the speech signal is corrupted by severe acoustical disturbances [2]. Since designing an algorithm that tries to cover all possible interferences and user scenarios is often impractical, finding the proper parameter values in an SE system for a given scenario is critical for real-world deployment. The system is often hand-tuned by experts and verified through subjective listening tests. However, the hand-tuning process is not only time-consuming but also error-prone since only a relatively small number of user scenarios can be covered.

Very little work has been done to formalize the tuning problem in SE systems, notably [3], due to the combinatorial nature of the problem and the related optimization criteria that rely on the fuzzy concept of *perceptually better quality* [4]. To get around the subjective and combinatorial nature of the design and tuning problem, locally optimal or near-optimal solutions are found by considering one component of the system at a time, and the concept of perceived quality is approximated by measures that are easy to describe mathematically, e.g., the mean squared error (MSE) or maximum likelihood (ML) [5]. However, it is well known that these types of measures, as well as the assumptions behind them, are hardly related to the auditory system [6], making the tuned solution suboptimal. Several methods have been proposed to objectively measure the perceived quality of speech signals, e.g., [7] and references therein. The mean opinion score (MOS) is the current standardized measure which compares a high quality fixed reference to its degraded version and ranks the result from "inaudible" to "very annoying" on a five-point scale [8]. This score can be calculated using automated techniques that mimic the human hearing process [9]. The most commonly used method is the Perceptual Evaluation of Speech Quality (PESQ) [10], but its scope is limited to speech codecs evaluation. A new model called Perceptual Objective Listening Quality Assessment (POLQA) [11] addresses many of the issues and limitations of PESQ and produces reliable scores for evaluating SE algorithms.

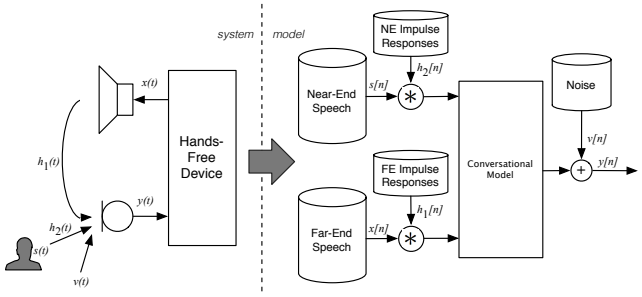


**Fig. 1**. Model of the loudspeaker-microphone configuration.

Besides the optimization criteria, constructing a comprehensive database that covers all possible scenarios is also essential to developing an effective SE algorithm, and recent works have focused on providing a common framework to test and evaluate SE algorithms, e.g., for noise suppression [12] or dereverberation [13]. However, to the authors' knowledge, there is currently no database for evaluating SE algorithms in full-duplex communication, which is the target of our system. Thus, a full-duplex communication database is often "handmade" but tailors to only a few scenarios.

In this work, we propose a formal procedure for tuning the parameters of an SE system for hands-free devices. The system comprises of an acoustic echo canceler (AEC), a residual echo power estimator (RPE), a noise power estimator (NPE), a residual echo suppressor (RES), and a noise suppressor (NS). The tuning problem is casted as an optimization problem where the cost function is a perceptual objective measure and the optimization variables are the parameters of the SE chain, and a genetic algorithm is used to determine the global solution. For this purpose, a large multi-condition database is automatically generated by considering the characteristics of human conversational speech. The database encompasses various key factors including room impulse responses (RIRs), noise types, speakers, echo return losses, and signal-to-noise ratios (SNRs), to model a real full-duplex communication as shown in Figure 1. We then compare different objective perceptual measures as optimization criteria and perform a subjective listening test on the different outputs obtained.

## 2. SPEECH ENHANCEMENT ALGORITHM

Let $y[n]$ be the near-end microphone signal, which consists of the near-end speech $s[n]$ and noise $v[n]$ mixed with the acoustic echo $d[n] = h[n] * x[n]$, where $h[n]$ is the impulse response of the system, $x[n]$ is the far-end reference signal, and $*$ is the convolution operator. The overall block diagram of the speech enhancement algorithm is shown in Figure 2. The AEC subtracts the linear part of the echo $d[n]$ while the RES/NS suppresses the nonlinear part of the residual echo $b[n] = d[n] - \hat{d}[n]$ and noise $v[n]$.

### 2.1. Robust Acoustic Echo Canceler

Since strong near-end interference may corrupt the error signal of the AEC and cause the adaptive filter to diverge, the robust acoustic echo canceler system [14–16] is used, where a error recovery nonlinearity (ERN) allows for continuous updating. To reduce the delay of the frequency-domain adaptive filter [17], the multi-delay adaptive filter structure [18] is used. The tuning parameters for the AEC consist of the number of partitioned blocks $M_{\text{AEC}}$, the number of iterations $N_{\text{AEC}}$, the step-size $\mu_{\text{AEC}}$, and the smoothing factor $\alpha_{\text{AEC}}$ for the power spectral density estimation.

### 2.2. Residual Echo Power Estimator

A coherence based method similar to [19, 20] is used for the RPE. The residual echo is modeled as (omitting the frame index $m$ whenever necessary for simplicity) $B_k = \Delta\mathbf{H}_k^{\mathrm{T}}\mathbf{X}_k$, where $\Delta\mathbf{H}_k = [\Delta H_k[0], \ldots, \Delta H_k[M_{\text{RPE}}-1]]^{\mathbf{T}}$ (system distance in the STFT domain) and $\mathbf{X}_k = [X_k[m], \ldots, X_k[m-M_{\text{RPE}}+1]]^{\mathbf{T}}$ for the $k^{\text{th}}$ frequency bin. The system distance can be estimated using a minimum mean-square error (MMSE) approach [20]:

$$\Delta\hat{\mathbf{H}}_k = \mathrm{E}\{\mathbf{X}_k^*\mathbf{X}_k^{\mathrm{T}}\}^{-1}\mathrm{E}\{\mathbf{X}_k^* B_k\} \equiv \Phi_{\mathbf{XX}}^{-1}[k]\Phi_{\mathbf{X}B}[k]. \quad (1)$$

Using only the diagonal terms of the autocorrelation matrix $\Phi_{\mathbf{XX}}$ and the error signal $E$ in place of the true residual echo $B$, the residual echo power is estimated by

$$\lambda_B[k] = \left|\frac{\hat{\Phi}_{\mathbf{X}E}^{\mathrm{T}}[k]\mathbf{X}_k}{\hat{\Phi}_{XX}[k]}\right|^2, \quad (2)$$

where

$$\hat{\Phi}_{\mathbf{X}E}[k,m] = \alpha_{\text{RPE}}\hat{\Phi}_{\mathbf{X}E}[k,m-1] + (1-\alpha_{\text{RPE}})\mathbf{X}_k^* E_k, \quad (3)$$

$$\hat{\Phi}_{XX}[k,m] = \alpha_{\text{RPE}}\hat{\Phi}_{XX}[k,m-1] + (1-\alpha_{\text{RPE}})|X_k|^2. \quad (4)$$

The tuning parameters for RPE consist of the number of past frames $M_{\text{RPE}}$ and the smoothing factor $\alpha_{\text{RPE}}$.

### 2.3. Noise Power Estimator

The low complexity MMSE noise power estimator [21] that implicitly accounts for the speech presence probability (SPP) is used for the NPE. The MMSE estimation of a noisy periodogram under speech presence uncertainty results in

$$\mathrm{E}\{\lambda_V[k]|E_k\} = P(H_1|E_k)\lambda_V[k] + P(H_0|E_k)|E_k|^2, \quad (5)$$

where the *a posteriori* SPP is calculated by

$$P(H_1|E_k) = \left[1 + (1+\xi_{H_1})\exp\left(-\frac{|E_k|^2}{\lambda_V[k,m-1]}\frac{\xi_{H_1}}{1+\xi_{H_1}}\right)\right]^{-1}. \quad (6)$$

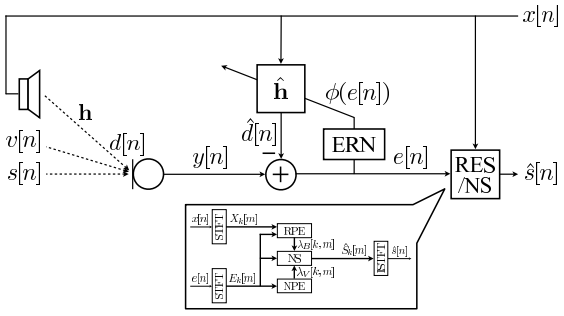


**Fig. 2**. Block diagram of the speech enhancement system.

The noise power spectral density is then updated by

$$\lambda_V[k,m] = \alpha_{\text{NPE}}\lambda_V[k,m-1] + (1-\alpha_{\text{NPE}})\mathrm{E}\{\lambda_V[k]|E_k\}. \quad (7)$$

To avoid stagnation due to an underestimated noise power, a smoothing is performed

$$\bar{P} = \alpha_P\bar{P} + (1-\alpha_P)P(H_1|E_k), \quad (8)$$

and the following ad-hoc procedure is used for the update:

$$P(H_1|E_k) = \begin{cases}\min\{P(H_1|E_k), P_{\text{TH}}\}, & \bar{P} > P_{\text{TH}},\\ P(H_1|E_k), & \text{otherwise.}\end{cases} \quad (9)$$

The tuning parameters for the NPE consist of the fixed *a priori* SNR $\xi_{H_1}$, the threshold $P_{\text{TH}}$, and the smoothing factors $\alpha_P$ and $\alpha_{\text{NPE}}$.

### 2.4. Noise Suppressor

The Ephraim and Malah log-spectral amplitude (LSA) MMSE estimator [22] is used for the NS:

$$G_k^{\text{LSA}} = \frac{\xi_k}{1+\xi_k}\exp\left(\frac{1}{2}\int_{\frac{\xi_k\gamma_k}{1+\xi_k}}^{\infty}\frac{e^{-t}}{t}\,\mathrm{d}t\right), \quad (10)$$

where the *a priori* SNR $\xi_k$ and the *a posteriori* SNR $\gamma_k$ are

$$\xi_k = \frac{\mathrm{E}\{|S_k|^2\}}{\mathrm{E}\{|V_k|^2\} + \mathrm{E}\{|B_k|^2\}} = \frac{\lambda_S[k]}{\lambda_V[k] + \lambda_B[k]}, \quad (11)$$

$$\gamma_k = \frac{\mathrm{E}\{|E_k|^2\}}{\mathrm{E}\{|V_k|^2\} + \mathrm{E}\{|B_k|^2\}} = \frac{\lambda_E[k]}{\lambda_V[k] + \lambda_B[k]}. \quad (12)$$

The estimation of the *a priori* SNR is done using the decision-directed (DD) approach [23]:

$$\xi_k^{\text{DD}} = \alpha_{\text{DD}}\frac{|\hat{S}_k[m-1]|^2}{\lambda_V[k,m]+\lambda_B[k,m]} + (1-\alpha_{\text{DD}})\max\{\gamma_k - 1, 0\}. \quad (13)$$

To further reduce the musical noise, the suppression gain is limited to a certain minimum value $G_{\text{min}}$:

$$\hat{S}_k = [(1-G_{\text{min}})G_k^{\text{LSA}} + G_{\text{min}}]E_k. \quad (14)$$

The tuning parameters of the NS consist of the smoothing factor for the SNR estimator $\alpha_{\text{DD}}$ and the minimum suppression gain $G_{\text{min}}$.

## 3. TUNING AS AN OPTIMIZATION PROBLEM

The tuning problem can be easily formulated as a general optimization problem [4], where the objective function to *maximize* is the speech quality, or MOS, produced by the SE system. Since most measures are full-referenced, we calculate the difference in MOS as

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

We can reasonably assume that inequality constraint functions are linear and univariate. Thus the constraints simplify to determining the lower and upper bounds for the components of the solution vector, and our optimization problem becomes:

$$\begin{aligned} &\text{maximize} \quad \Delta\text{MOS}\left(\hat{s}[n, \mathbf{p}], y[n]\right) \\ &\text{subject to} \quad \mathbf{U} \leqslant \mathbf{p} \leqslant \mathbf{L}. \end{aligned} \tag{15}$$

where $\mathbf{p}$ is the vector of parameters that needs tuning, $\hat{s}[n, \mathbf{p}]$ is the SE system output obtained with $\mathbf{p}$, and $\mathbf{L}$ and $\mathbf{U}$ represent, respectively, the lower and upper bounds in each element of $\mathbf{p}$. While not strictly necessary, explicitly defining these bounds in our formulation allows us to obtain faster and more reliable solutions.

Since the objective function is nonlinear and not known to be convex, there is no effective method for solving (15), e.g., performing a brute force search with as few as a dozen variables can be intractable. The general nonlinear programming problem can be solved by several approaches, each of which involves some compromises [24]. The so-called genetic algorithm has been successfully applied to this type of non-convex mixed-integer optimization [25].

The basic idea is to apply genetic operators, such as *mutation* and *crossover*, to evolve a set of $M$ solutions, or *population*, $\mathbf{\Pi}^{(k)} = \{\mathbf{p}_m^{(k)}, m = 1, \ldots, M\}$ in order to find the solution that maximizes the cost function. This procedure begins with a randomly chosen population $\mathbf{\Pi}^{(0)}$ in the space of the feasible values $[\mathbf{L}, \mathbf{U}]$ and it is repeated until a halting criterion is reached after $K$ iterations. The set of parameters $\mathbf{p}_m^{(K)} \in \mathbf{\Pi}^{(K)}$ that maximizes the cost function will be our estimate:

$$\hat{\mathbf{p}} = \underset{\mathbf{p}_m^{(K)} \in \mathbf{\Pi}^{(K)}}{\arg\max}\ \Delta\text{MOS}\left(\hat{s}[n, \mathbf{p}_m^{(K)}], y[n]\right) \tag{16}$$

## 4. DATABASE GENERATION

The modeling of human conversational speech and the so-called conversational events, such as talk-spurt, pause, mutual silence, and double-talk, is fundamental to characterizing realistic scenarios in full-duplex communication. In particular, the studies done in [26] and [27] had a direct impact on the method for generating artificial conversational speech presented in [28]. However, this method is rather simplistic and relies on hand-coded expert knowledge [29], which is not easily transferable to the automatic generation of a large conversational speech database.

Several new methodologies have been proposed to model the turn-taking behaviors, e.g., [30] and references therein. However, these methodologies are focused on human-machine turn-taking with very little mutual social interaction. We therefore focus on older studies on human-human conversations like [27]. In particular, we propose a flexible model of conversational behavior using a 4-state Markov chain model, where the states correspond to, respectively, mutual silence (MS), near-end (NE) talk, far-end (FE) talk, and double-talk (DT), and define all the possible combinations of the components in $y[n]$.

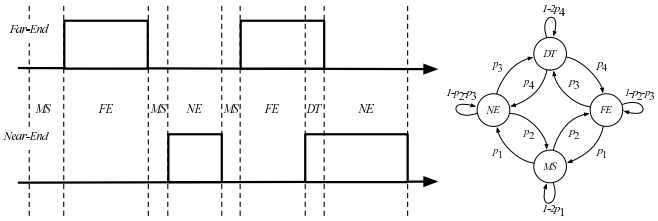


**Fig. 3**. Conversational sequence and its Markov chain model.

The Markov chain is uniquely described by its transition matrix $\mathbf{T}$ to model the generation model in [28] and the related distributions of the conversational events. According to the distribution of the single talk duration, $T_{ST}$, the double talk duration, $T_{DT}$, and the mutual silence duration, $T_{MS}$, presented in [28], we are able to use a Markov chain Monte Carlo (MCMC) sampling algorithm [31] to find the transition matrix $\mathbf{T}$ of the 4-state Markov chain. Given that the transition between active NE and active FE and the transition between MS and DT are not allowed, and that the transition probabilities of going from MS to NE and MS to FE are equivalent [28], the Markov chain is uniquely represented by only four parameters:

$$\mathbf{T} = \begin{bmatrix} 1-2p_1 & p_1 & p_1 & 0 \\ p_2 & 1-p_3-p_2 & 0 & p_3 \\ p_2 & 0 & 1-p_3-p_2 & p_3 \\ 0 & p_4 & p_4 & 1-2p_4 \end{bmatrix}. \tag{17}$$

This makes it very easy to modify and fit different types of conversation scenarios with different levels of interactivity [32]. An example of a sequence of conversational speech and its Markov chain model is shown in Figure 3.

## 5. EXPERIMENTAL ANALYSIS

In the experimental evaluation, the optimization framework presented in Section 3 was used to tune and evaluate the SE algorithm presented in Section 2. Here we provide details for the method proposed and the results obtained.

### 5.1. Setup

The speech databases were generated using the ITU-T P-Series test signals [33]. This set includes 16 recorded sentences in each of 20 languages and sentences recorded in an anechoic environment, sampled at 16 kHz. From these, we generated two single-channel signals, NE and FE, with continuous activity (i.e., without pauses). The total duration of the speech is about one hour per channel. The NE and FE speech segments were generated using the Markov chain presented in Section 4 with $p_1 = 0.04$, $p_2 = 0.03$, $p_3 = 0.05$, and $p_4 = 0.25$, generating the same statistical behavior of conversational events as specified in [28].

A noise database comprised of babble (e.g., airport, cafeteria, exhibition, and restaurant) noise, white and pink noise, impulsive noise (e.g., hammering), airplane cabin noise, car noise from a variety of car models, and street noise was used. The RIRs were calculated in office environments using the Audio Precision APx525 log-swept chirp signal through the *Beats Pill™* portable speaker and truncated to the desired length ($f_s$= 48 kHz, resampled at 16 kHz). A set of 10 RIRs was then chosen with average reverberation time, $\text{RT}_{60}$, of 0.28 s [34].

In order to generate the NE and FE segments, the starting and ending points were chosen randomly within the NE and FE channels.

We generated 1000 segments with lengths between 6 to 8 s, ideal for objective quality measures [10, 11]. The two segments were then normalized to -26 dBov to avoid clipping, following the ITU-T Recommendation P.835 [35], and convolved with their respective RIR with normalized unitary energy. The microphone signal was created as follows. The NE signal was mixed with the FE signal at signal-to-echo ratios (SERs) uniformly distributed between -30 and 5 dB. The scaling was done by calculating the energy of the signals according to [36]. The noise was then mixed at an SNR uniformly distributed between -5 to 10 dB, according to the noise and the mixed speech signal energies [12].

Considering the SE algorithm presented in Section 2 and the problem in (15), we define the parameter vector as

$$\mathbf{p} = \{M_{\text{AEC}}, N_{\text{AEC}}, \mu_{\text{AEC}}, \alpha_{\text{AEC}}, M_{\text{RPE}}, \alpha_{\text{RPE}}, \xi_{H_1}, P_{\text{TH}}, \alpha_P, \alpha_{\text{NPE}}, \alpha_{\text{DD}}, G_{\min}\}, \quad (18)$$

and empirically determine reasonable upper and lower bounds for each variable. The genetic algorithm had a population of $M = 20$ possible candidates, and the best $N = 4$ were migrated to the next generation. These values were chose empirically balancing complexity and accuracy of the results. Of the remaining sets, half went through crossover and half went through mutation (uniform mutation was chosen). The perceptual objective quality measure used was the average $\Delta$MOS, as obtained through PESQ [10], POLQA [11], and the recently introduced ViSQOL [37, 38]. We included the manually tuned system, where the parameters were selected during the algorithmic design phase as a reference, and obtained four sets of parameters: $\mathbf{p}_{\text{POLQA}}$, $\mathbf{p}_{\text{PESQ}}$, and $\mathbf{p}_{\text{ViSQOL}}$, and $\mathbf{p}_{\text{MANUAL}}$. For comparison, we also optimized the SE system over four traditional objective measures, averaged over the evaluation set, that do not account for perception: log-spectral distortion (LSD), true echo return loss enhancement (tERLE), MSE, and a combined measure where the AEC block is optimized first using tERLE, and the RPE, NPE, and NS blocks are optimized with LSD (with fixed AEC parameters). The following sets were obtained with proposed optimization method: $\mathbf{p}_{\text{LSD}}$, $\mathbf{p}_{\text{tERLE}}$, $\mathbf{p}_{\text{MSE}}$, and $\mathbf{p}_{\text{tERLE+LSD}}$.

### 5.2. Results

We divided the database into two parts, where 80% was used to estimate the parameters and 20% was used for testing. Table 1 shows the $\Delta$MOS calculated using PESQ, POLQA, ViSQOL, and various traditional objective measures. The results show a net improvement in MOS over the manually tuned method, which in turn outperforms all the traditional objective measures. This proves that, in general, a trained ear is much better at determining proper values for the various parameters than using only the traditional objective measures, even if the tuning is done on a limited set. However, the use of perceptual objective measures for large-scale optimization greatly improves the performance of the SE algorithm over a much larger dataset. $\Delta\text{MOS}_{\text{POLQA}}$, arguably the most reliable measure for SE performance evaluation, 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.

A subjective evaluation was performed through the MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) test [39]. We compared the manually tuned configuration $\mathbf{p}_{\text{MANUAL}}$ with the two configurations obtained with standardized ITU-T tools, $\mathbf{p}_{\text{POLQA}}$ and $\mathbf{p}_{\text{PESQ}}$. The anchors were chosen as a 3.5 kHz low-pass filtered version (LP3.5) of the reference signal for scaling, as specified in [39], and the unprocessed speech, to represent the worst-case scenario in the listening evaluation. A pool of eleven expert listeners, familiar in detecting small impairments, and seven naive listeners was chosen. The test was performed using six speech excerpts randomly selected from the testing database. The results shown in Figure 4 are in line with the objective analysis. In particular, the confidence interval of the POLQA score only minimally overlaps with other scores, showing a significant statistical difference. The high variance of the LP3.5 and manually tuned scores is explained by the observed bimodality of the distribution of these scores, with a good percentage of the subjects preferring the bandlimitedness of the anchor over the manually tuned enhanced speech. Nonetheless, all subjects consistently preferred the POLQA-based tuning.

**Table 1**. 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 |

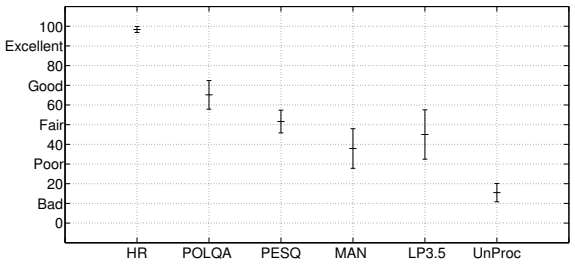


**Fig. 4**. Results of the MUSHRA listening test comparing three different tuning strategies: POLQA-based, PESQ-based, and manual tuning.

## 6. CONCLUSIONS

We have presented a methodology to tune the parameters of a speech enhancement system for full-duplex communication. The values of the parameters are often chosen empirically in the development stage of the algorithmic design and are most likely suboptimal. We have shown that optimizing over an objective criterion that embeds aspects of human perception works well in determining better solutions to the tuning problem. The MUSHRA test shows a fairly significant preference over the manually tuned system. Furthermore, using standardized objective quality measures like PESQ and POLQA, we see a net increase in MOS, usually not easily obtained without significant algorithmic changes. In order to perform the large scale optimization, we implemented a method to construct a database which creates realistic full-duplex communication scenarios. The methodology presented is a first step toward a more elegant way to handle the tuning problem, helping the deployment process, guiding the algorithm development, and highlighting shortcomings of the system.

## 7. REFERENCES


[1] P. C. Loizou, *Speech Enhancement: Theory and Practice*, CRC press, 2007.

[2] J. Benesty, M. Sondhi, and Y. Huang, *Springer Handbook of Speech Processing*, Springer, 2008.

[3] 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.

[4] D. Giacobello, J. Atkins, J. Wung, and R. Prabhu, "Results on Automated Tuning of a Voice Quality Enhancement System Using Objective Quality Measures," in *Proc. Audio Engineering Society Convention*, 2013.

[5] I. Tashev and M. Slaney, "Data Driven Suppression Rule for Speech Enhancement," *Proc. Information Theory and Applications Workshop*, pp. 1–6, 2013.

[6] M. G. Christensen and S. H. Jensen, "On Perceptual Distortion Minimization and Nonlinear Least-Squares Frequency Estimation," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 99–109, 2006.

[7] Y. Hu and P. C. Loizou, "Evaluation of Objective Quality Measures for Speech Enhancement," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, 2008.

[8] *Methods for Subjective Determination of Transmission Quality*, ITU-T Rec. P.800, 1996.

[9] S. Moller, W.-Y. Chan, N. Côté, T. H. Falk, A. Raake, and M. Waltermann, "Speech Quality Estimation: Models and Trends," *IEEE Signal Processing Magazine*, vol. 28, no. 6, pp. 18–28, 2011.

[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] Y. Hu and P. C. Loizou, "Subjective Comparison and Evaluation of Speech Enhancement Algorithms," *Speech Communication*, vol. 49, no. 7, pp. 588–601, 2007.

[13] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, A. Sehr, W. Kellermann, and R. Maas, "The REVERB Challenge: A Common Evaluation Framework for Dereverberation and Recognition of Reverberant Speech," *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 1–4, 2013.

[14] T. S. Wada and B.-H. Juang, "Acoustic Echo Cancellation Based on Independent Component Analysis and Integrated Residual Echo Enhancement," *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 205–208, 2009.

[15] 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.

[16] 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.

[17] J. J. Shynk, "Frequency-Domain and Multirate aAdaptive Filtering," *IEEE Signal Processing Magazine*, vol. 9, no. 1, pp. 14–37, 1992.

[18] J. S. Soo and K. K. Pang, "Multidelay Block Frequency Domain Adaptive Filter," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 38, no. 2, pp. 373–376, 1990.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] ——, "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.

[24] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[25] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, John Wiley & Sons, 2012.

[26] P. T. Brady, "A Statistical Analysis of On-Off Patterns in 16 Conversations," *The Bell System Technical Journal*, vol. 47, pp. 73–91, 1968.

[27] ——, "A Model for Generating On-Off Speech Patterns in Two-Way Conversation," *The Bell System Technical Journal*, vol. 48, pp. 2445–2472, 1969.

[28] *Artificial Conversational Speech*, ITU-T Rec. P. 59, 1993.

[29] F. Kronlid, "Turn Taking for Artificial Conversational Agents," *Proc. International Conference on Cooperative Information Agents*, pp. 81–95, 2006.

[30] A. Raux and M. Eskenazi, "A Finite-State Turn-Taking Model for Spoken Dialog Systems," *Proc. North American Chapter of the Association for Computational Linguistics - Human Language Technologies*, pp. 629–637, 2009.

[31] C. Andrieu, N. De Freitas, A. Doucet, and M. I. Jordan, "An Introduction to MCMC for Machine Learning," *Machine learning*, vol. 50, pp. 5–43, 2003.

[32] F. Hammer, P. Reichl, and A. Raake, "Elements of Interactivity in Telephone Conversations," *Proc. International Conference on Spoken Language Processing*, pp. 1741–1744, 2004.

[33] *Telephone Transmission Quality, Telephone Installations, Local Line Networks*, ITU-T P. Series. Available: http://www.itu.int/net/ITU-T/sigdb/genaudio/Pseries.htm

[34] M. R. Schroeder, "New Method of Measuring Reverberation Time," *The Journal of the Acoustical Society of America*, vol. 37, pp. 409–412, 1965.

[35] *Subjective Test Methodology for Evaluating Speech Communication Systems that Include Noise Suppression Algorithms*, ITU-T Rec. P. 835, 2003.

[36] *Objective Measurement of Active Speech Level*, ITU-T Rec. P.56, 1993.

[37] 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.

[38] ——, "Robustness of Speech Quality Metrics to Background Noise and Network Degradations: Comparing ViSQOL, PESQ and POLQA," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3697–3701, 2013.

[39] *Method for the Subjective Assessment of Intermediate Sound Quality*, ITU-T Rec. 1534-1, 2001.