Audio Engineering Society
Convention e-Brief 114
Presented at the 135th Convention
2013 October 17–20 New York, USA

Results on Automated Tuning of a Voice Quality Enhancement System Using Objective Quality Measures

Daniele Giacobello, Joshua Atkins, Jason Wung, and Raghavendra Prabhu

Beats by Dr. Dre, 1601 Cloverfield Blvd., Suite 5000N, Santa Monica, CA 90404, USA

ABSTRACT

In this work, we present a formal procedure for the automated tuning of various parameters in a voice quality enhancement system. Firstly, we formulate the problem of tuning as a large-scale nonlinear programming problem. Secondly, we evaluate the performance of perceptual objective quality measures as optimization criteria for our tuning problem. We then perform a subjective quality assessment to compare the output of a voice quality enhancer obtained with parameters calculated with these different criteria and also with those obtained through a conventional approach of tuning by expert listening. The results show that this automated methodology performs well in finding reasonable solutions for the tuning problem, potentially saving time and resources over manual evaluation and tuning.

1. INTRODUCTION

Voice quality enhancement (VQE) algorithms are integral to many diverse speech communication devices such as hearing aids and cellular phones [1]. In particular, VQE algorithms are crucial in extending the usage of these devices to scenarios with severe acoustical disturbances [2].

A central aspect of the VQE algorithm design is to properly tune all the different components in order to handle most, if not all, possible application scenarios. Tuning is generally recognized throughout the literature as a significant bottleneck when developing any type of system given the intrinsic combinatorial nature of the problem [3]. This becomes even harder for VQE algorithms, and audio systems in general, where the optimization criterion relates to the fuzzy concept of *perceptually better quality* [4]. This makes subjective listening tests and trained ears still considered as the most reliable way of measuring the quality of an audio system. However, this can be a very time consuming process, often taking much longer than the algorithm design and implementation phases. Furthermore, the human component in the tuning process makes it error-prone and bound to cover only a relatively small number of application scenarios.

The classic approach to get around the subjective nature of the design and tuning problem is to approximate the concept of perceived quality with metrics easier to describe mathematically such as the mean squared error (MSE) [5]. These approximations often poorly relate to the auditory system [6], making the tuning solution highly suboptimal.

The standard metric for measuring the perceived quality of speech signals is the mean opinion score (MOS) [7] which ranks the degradation of a VQE system compared to a high quality fixed reference from "inaudible" to "very annoying" on a five-point scale. This score can be calculated using automated techniques that mimic the human hearing process [8]. The most commonly used method is the Perceptual Evaluation of the Speech Quality (PESQ) [9]. However, given its limited scope to speech codecs evaluation, a new model called Perceptual Objective Listening Quality Assessment (POLQA) [10] was developed. POLQA addresses many of the issues and limitations of PESQ and is meant to produce reliable scores for VQE tuning.

In this paper, we provide initial results on the automation of the tuning process of a VQE algorithm using objective perceptual measures of speech quality. While computationally expensive

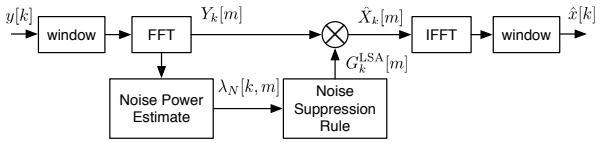

Fig. 1: A block diagram of a VQE system for noise reduction.

and hard to analyze, the MOS produced by these measures correlates with human perception much better than conventional approaches such as MSE. To the best of our knowledge, the attempt to use these measures for the design and tuning process of VQE algorithms was only found in [4].

2. NOISE SUPPRESSION

A number of algorithms have been proposed for noise reduction throughout the years. These algorithms can be divided into three main categories: spectral subtractive algorithms, statistical model-based algorithms, and subspace algorithms [1]. We will focus our attention to the statistical model-based category which includes some of the most popular algorithms.

Let $y[k] = y(kT)$ represent values from a band-limited and time-limited speech signal, uniformly sampled at $T = 1/f_s$, where f_s is the sampling frequency. The corrupted sequence of speech can be represented by the additive model: $y[k] = x[k] + n[k]$, where $y[k]$ is the observed signal, $n[k]$ is the additive noise, and $x[k]$ is the desired clean speech. The goal of the noise suppressor is to form an estimate, $\hat{x}[k]$, of $x[k]$ based on the observed signal $y[k]$. For real-time implementation, the noisy input signal is usually divided into subsequent overlapping frames of short duration. The signal frames are then transformed into the frequency domain through the discrete Fourier transform (DFT). This is also called the analysis-modification-synthesis (AMS) scheme that wraps around most of the currently deployed VQE algorithms. Its objective is twofold. Firstly, AMS divides the signal into manageable frames within which the underlying statistical properties can be assumed to be invariant. Secondly, by applying the DFT, it delivers approximately uncorrelated transform coefficients [11]. Taking a K-point DFT of the m^{th} windowed frame yields K complex frequency components:

$$Y_k[m] = X_k[m] + N_k[m], \quad k = 0, \ldots, K-1.$$

The noise suppression problem is then to retrieve $X_k[m]$ based on $Y_k[m]$. Two components are generally required, a noise power estimator and a suppression rule based on this estimate. In the remainder of the paper we will omit the frame index, m, wherever possible. A block diagram of the VQE system considered is shown in Fig. 1.

2.1. Noise Suppression Rule

The idea of noise reduction is to apply a gain on the noisy speech to obtain an estimate of the clean speech signal: $\hat{X}_k = G_k Y_k$. The simplest approach to noise reduction is to estimate the clean speech DFT coefficients using a linear minimum MSE (MMSE) estimator [11], often referred to as the frequency domain Wiener filter [12]. The Wiener filter assumes stationary and statistically uncorrelated speech and noise signals, i.e., $\mathrm{E}\{X_k N_k^*\} = 0$, and is given by $G_k^{\mathrm{W}} = \xi_k/(1+\xi_k)$, where $\xi_k = \mathrm{E}\{|X_k|^2\}/\mathrm{E}\{|N_k|^2\} = \lambda_X[k]/\lambda_N[k]$ is the so-called *a priori* signal-to-noise ratio (SNR) that is usually estimated by

$$\xi_k = \beta \frac{|\hat{X}_k[m-1]|^2}{\lambda_N[k,m]} + (1-\beta)\max\left\{\frac{|Y_k[m]|^2}{\lambda_N[k,m]} - 1, 0\right\}. \quad (1)$$

Under similar assumptions, Ephraim and Malah derived a log-spectral amplitude (LSA) MMSE estimator [13]:

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

Perceptually speaking, (2) has much less musical noise than the Weiner filter [11], thus making it particularly popular throughout speech enhancement literature [1]. To further reduce the musical noise, the suppression gain is limited in practice to a certain minimum value:

$$\hat{X}_k = ((1-m_G)G_k^{\mathrm{LSA}} + m_G)Y_k. \quad (3)$$

2.2. Noise Power Estimation

The presented suppression rules are heavily dependent on a proper estimate of the spectral noise power, usually obtained in combination with a voice activity detector. A newly proposed algorithm [14] allows for implicitly accounting for the speech presence probability (SPP). The algorithm is briefly described as follows (see [14] for more detail).

The MMSE estimation of a noisy periodogram under speech presence uncertainty results in

$$\mathrm{E}\{\lambda_N[k]\big|Y_k\} = P(H_1\big|Y_k)\hat{\lambda}_N[k] + (1 - P(H_1\big|Y_k))|Y_k|^2, \quad (4)$$

where the *a posteriori* SPP is calculated by

$$P(H_1\big|Y_k) = \left[1 + (1+\xi_{H_1})\exp\left(-\frac{|Y_k|^2}{\hat{\lambda}_N[k]}\frac{\xi_{H_1}}{1+\xi_{H_1}}\right)\right]^{-1}, \quad (5)$$

under the assumption of uniform priors, i.e., $P(H_0) = P(H_1) = 0.5$. $\hat{\lambda}_N[k] = \lambda_N[k, m-1]$ is the previous-frame estimate of the noise power spectral density (PSD) that is updated by

$$\lambda_N[k,m] = \alpha_{\mathrm{PSD}}\lambda_N[k,m-1] + (1-\alpha_{\mathrm{PSD}})\mathrm{E}\{\lambda_N[k]|^2\big|Y_k\}. \quad (6)$$

In [14], the author chose a fixed *a priori* SNR ξ_{H_1} by minimizing the sum of false-alarm and missed-hit probabilities of the speech presence estimator.

In order to avoid stagnation due to an underestimated noise power, a check is performed to see if the *a posteriori* SPP has been close to one for too long. A smoothing is performed:

$$\bar{P} = \alpha_P \bar{P} + (1-\alpha_P) P(H_1|Y_k), \tag{7}$$

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

$$P(H_1|Y_k) = \begin{cases} \min\{P_{\text{TH}}, P(H_1|Y_k)\}, & \text{if } \bar{P} > P_{\text{TH}}, \\ P(H_1|Y_k), & \text{otherwise.} \end{cases} \tag{8}$$

3. TUNING AS AN OPTIMIZATION PROBLEM

A general mathematical optimization problem has the form

$$\begin{aligned} &\text{minimize} \quad D(\mathbf{p}) \\ &\text{subject to} \quad U_i \le f_i(\mathbf{p}) \le L_i, \quad i = 1,\dots,C, \end{aligned} \tag{9}$$

where $\mathbf{p} = \{p_1, p_2, \dots, p_N\} \in \mathbb{R}^N$ is a vector of the optimization variables, $D(\cdot) : \mathbb{R}^N \to \mathbb{R}$ is an objective function, $f_i(\cdot) : \mathbb{R}^N \to \mathbb{R}$, $i = 1,\dots,C$ are inequality constraint functions, and $\{U_i, L_i\}$ are limits (bounds) for the constraint functions.

The tuning problem can be easily formulated as the above optimization problem, where an objective function ΔMOS, the increase in MOS produced by the VQE, is calculated from

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

Assuming that the inequality constraint functions are linear and univariate, the constraints simplify to lower and upper bounds of the solution vector, and the optimization problem becomes

$$\begin{aligned} &\text{maximize} \quad \Delta\text{MOS}\left(\hat{x}[k]|_{\mathbf{p}}, y[k]\right) \\ &\text{subject to} \quad \mathbf{U} \le \mathbf{p} \le \mathbf{L}, \end{aligned} \tag{10}$$

where $\mathbf{p}$ is now a vector of parameters to be tuned, $\hat{x}[k]|_{\mathbf{p}}$ is the VQE output obtained with these parameters, and $\mathbf{L}$ and $\mathbf{U}$ represent, respectively, the lower and upper bounds of the values for each variable. While not strictly necessary, explicitly defining these bounds in our formulation allows us to obtain fast and reliable solutions. Since the objective function is neither linear nor convex, there is no effective method for solving (10). Performing a brute force search with as few as ten variables can be extremely challenging, while problems with a few hundreds of variables can be intractable. Therefore, methods to solve the general nonlinear programming problem utilize several different approaches, each of which involves some compromises [15].

	β	m_G	ξ_{H_1}	α_P	P_{TH}	α_{PSD}
U	0.98	0.050	63.24	0.90	0.99	0.90
L	0.80	0.001	10.00	0.45	0.90	0.65
$\mathbf{p}_{\text{HUMAN}}$	0.98	0.010	31.62	0.50	0.99	0.80
$\mathbf{p}_{\text{LSD}}$	0.87	0.003	28.78	0.62	0.94	0.83
$\mathbf{p}_{\text{PESQ}}$	0.91	0.010	42.25	0.64	0.95	0.67
$\mathbf{p}_{\text{POLQA}}$	0.95	0.040	21.15	0.76	0.96	0.76

Table 1: Values resulting from the optimization methods for each parameter vector. The first two rows represent the bounds imposed on the parameter vector in the optimization process.

4. EXPERIMENTAL RESULTS

The parameters to be estimated are $\{\beta, m_G\}$ for the noise suppressor and $\{\xi_{H_1}, \alpha_P, P_{\text{TH}}, \alpha_{\text{PSD}}\}$ for the noise power estimator. Considering the problem in (10), we define

$$\mathbf{p} = \{\beta, m_G, \xi_{H_1}, \alpha_P, P_{\text{TH}}, \alpha_{\text{PSD}}\}.$$

To get around the combinatorial nature of the optimization problem in (10), we chose the so-called *genetic algorithm* [16] which can determine global solutions to problems that contain multiple maxima or minima.

The evaluation corpus was synthetically generated by mixing the ITU-T P-Series test signals [17], downsampled to 8 kHz, with a noise database composed of car, babble (airport, exhibition, etc.), fan, white, and pink noise. Noise and speech were mixed at SNRs ranging from -5 to 25 dB following the ITU-T Recommendation P. 835 [18] where the reference signal was always scaled to an ideal average active level of approximately -26 dBov to avoid clipping in the mixed signals. The AMS scheme was designed using a 128-sample (16 ms) Hamming window a 50% overlap.

The total length of all the combined audio signals amounts to about 1.5 hours, while the length of each audio signal is equally distributed between 8 to 16 seconds with roughly 50% of speech activity. We randomly picked 80% of the corpus for training and 20% for testing. For the tuning process presented in (10), we used PESQ and POLQA evaluated over the training corpus. For comparison purposes, we also optimized over the log-spectral distortion (LSD), known to provide a reasonable criterion for speech quality estimation [8], averaged over the known active speech regions. Through the automated tuning, we obtained three sets of parameters, $\mathbf{p}_{\text{POLQA}}$, $\mathbf{p}_{\text{PESQ}}$, and $\mathbf{p}_{\text{LSD}}$. $\mathbf{p}_{\text{HUMAN}}$ is the set of parameters tuned by expert listening over a limited set of the training corpus.

The values obtained through the optimization procedure for each method are shown in Table 1. The upper and lower bounds used in the optimization problem were determined empirically. A significant difference in the resulting values can be observed,

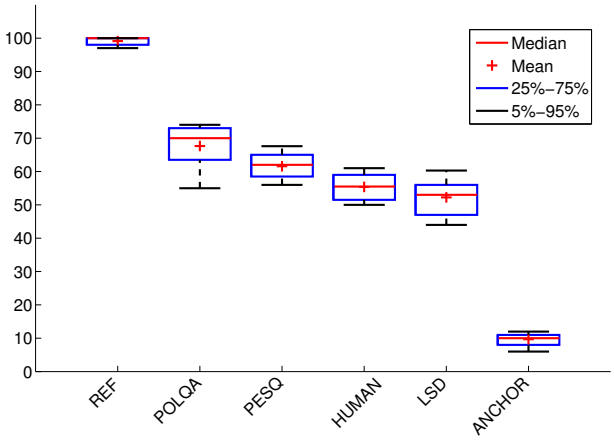

Fig. 2: Results of the MUSHRA listening test. MOS for different enhancement types averaged over all excerpts and all listeners. The boxes span the first and third quartile, the whiskers indicate the 95% confidence intervals.

especially compared to $\mathbf{p}_{\text{HUMAN}}$ where the values were chosen mostly based on what presented in the literature (e.g., the fixed *a priori* SNR ξ_{H_1} in [14]).

We then setup a MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) test to subjectively evaluate the perceptual quality of the four configurations considered. A pool of seven expert listeners, familiar in detecting small impairments, and five naive listeners was chosen. The test was performed using six speech clips randomly selected from the testing corpus. The anchor was created by low-pass filtering the clean signal at 1.75 kHz before performing the noise mixing (no enhancement in this case). The results are shown in Fig. 2.

Despite a significant overlap of the 95% confidence regions, speech obtained with $\mathbf{p}_{\text{POLQA}}$ and $\mathbf{p}_{\text{PESQ}}$ clearly achieves higher mean and median score of all the tuning methodologies. Furthermore, their median values do not fall within the interquartile range of the one another, thus suggesting that the set of tuning parameters obtained through POLQA offers better performance compared to the one obtained through PESQ. The interquartile ranges of the combined results obtained with $\mathbf{p}_{\text{POLQA}}$ and $\mathbf{p}_{\text{PESQ}}$ MUSHRA scores do not overlap with those obtained through $\mathbf{p}_{\text{LSD}}$ or $\mathbf{p}_{\text{HUMAN}}$, suggesting a significant difference between them. The median values of $\mathbf{p}_{\text{LSD}}$ and $\mathbf{p}_{\text{HUMAN}}$ fall within the interquartile range of one another, suggesting that there is no significant difference between them.

In terms of objective score, calculated via POLQA and PESQ, the VQE algorithm tuned with $\mathbf{p}_{\text{POLQA}}$ achieves roughly 0.2 increase in MOS compared to the hand tuned method. This is to be expected given that the optimization criteria are the MOS calculated with these algorithm.

5. CONCLUSIONS

We have provided some initial experimental results on the application of automated tuning for VQE algorithms using POLQA and PESQ, objective measures of perceptual speech quality that predict MOS. Optimizing over an objective criterion that embeds aspect of human perception seems to work reasonably well in determining better solutions to the tuning problem. In particular, a MUSHRA test showed a fairly significant preference over systems tuned either manually or with objective non-perceptual measure such as LSD.

6. 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] R. Freedman and G. J. Stuzin, "A Knowledge-Based Methodology For Tuning Analytical Models," *IEEE Trans. Syst., Man, Cybern.*, vol. 21, no. 2, pp. 347–358, 1991.

[4] I. Tashev, A. Lovitt, and A. Acero, "Unified Framework for Single Channel Speech Enhancement," in *Proc. IEEE PACRIM*, 2009, pp. 883–888.

[5] I. Tashev and M. Slaney, " Data Driven Suppression Rule for Speech Enhancement," in *Proc. ITA*, 2013.

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

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

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

[9] *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.

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

[11] R. C. Hendriks, T. Gerkmann, and J. Jensen, *DFT-Domain Based Single-Microphone Noise Reduction for Speech Enhancement: A Survey of the State of the Art*, Synthesis Lectures on Speech and Audio Processing, Morgan & Claypool, 2013.

[12] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*, MIT press, 1964.

[13] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator," *IEEE Trans. Acoust., Speech, Sig. Proc.*, vol. 33, no. 2, pp. 443–445, 1985.

[14] T. Gerkmann and R. Hendriks, "Unbiased MMSE-Based Noise Power Estimation With Low Complexity and Low Tracking Delay," *IEEE Audio, Speech, Lang. Proc.*, vol. 20, no. 4, pp. 1383–1393, 2012.

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

[16] D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley, 1989.

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

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