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

Daniele Giacobello, Joshua Atkins, Jason Wung, and Raghavendra Prabhu
Beats Electronics, LLC

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

Email: daniele.giacobello@beatsbydre.com

Motivation

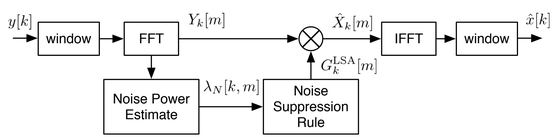
- Problems with tuning VQE algorithms:
  - necessity to tune over most possible application scenarios with very different acoustic disturbances (combinatorial problem),
  - optimization criterion is highly subjective (i.e., *perceived quality*).
- Proposed approach:
  - formalize the tuning process as a large-scale nonlinear optimization problem (1),
  - use of objective and reproducible measures as optimization criterion that mimic the human hearing process (e.g., PESQ (2) and POLQA (3)).

1 A Simple VQE: Noise Suppressor

We consider the following model:

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

Two components are generally required to estimate the clean speech $\hat{X}_k[m]$: a noise power estimator and a suppression rule based on this estimate.



A block diagram of a VQE system for noise reduction.

1.1 Noise Suppression Rule

- Apply a gain on the noisy speech to obtain an estimate of the clean speech signal: $\hat{X}_k = G_k Y_k$.
- A log-spectral amplitude (LSA) MMSE estimator is derived in (4):

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

where $G_k^{\text{W}} = \xi_k/(1+\xi_k)$ and $\xi_k = \mathrm{E}\{|X_k|^2\}/\mathrm{E}\{|N_k|^2\} = \lambda_X[k]/\lambda_N[k]$ is the so-called *a priori* SNR, estimated via the D-D approach:

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

- The suppression gain is limited to a minimum value:

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

1.2 Noise Power Estimation

- According to (5), the MMSE estimation of a noisy periodogram under speech presence uncertainty results in

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

where the *a posteriori* Speech Presence Probability (SPP) is calculated by

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

and the estimate of the noise PSD is

$$\hat{\lambda}_N[k,m] = \alpha_{\text{PSD}}\lambda_N[k,m-1] + (1-\alpha_{\text{PSD}})\mathrm{E}\{\lambda_N[k]|^2|Y_k\}.$$

- An ad-hoc procedure is performed to avoid stagnation of the *a posteriori* SPP:

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

where $\bar{P} = \alpha_P\bar{P} + (1-\alpha_P)P(H_1|Y_k)$.

2 Tuning as an Optimization Problem

- We want to optimize over the improvement in MOS produced by the VQE:

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

- The optimization problem is

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

  - $\mathbf{p}$ is the vector of parameters to be tuned.
  - $\hat{x}[k]|_{\mathbf{p}}$ is the VQE output obtained with $\mathbf{p}$.
  - $\mathbf{L}$ and $\mathbf{U}$ represent the constraints (lower and upper bounds) of the solution vector.

3 Experimental Setup

- **Evaluation corpus**
  - ITU-T P-Series speech test signals mixed with a noise database composed of car, babble, fan, white, and pink noise.
  - SNRs from -5 to 25 dB (reference: -26 dBov).
  - total length of all the combined audio signals about 1.5 hours (80% training).
  - length of each audio signal equally distributed between 8 to 16 s (50% of speech activity).
- **Optimization framework**
  - We optimize over the following parameters:

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

  - PESQ and POLQA were used to calculate the MOS. For comparison purposes, we also optimized over the LSD.
  - A *genetic algorithm* was chosen as effective in determining global solutions to nonlinear combinatorial problems (6).
  - The upper and lower bounds used in the optimization problem were determined empirically.
  - Output sets are $\mathbf{p}_{\text{POLQA}}$, $\mathbf{p}_{\text{PESQ}}$, and $\mathbf{p}_{\text{LSD}}$.
  - $\mathbf{p}_{\text{HUMAN}}$ is the vector tuned by expert listening over a limited set of the training corpus.
- **MUSHRA test**
  - 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.

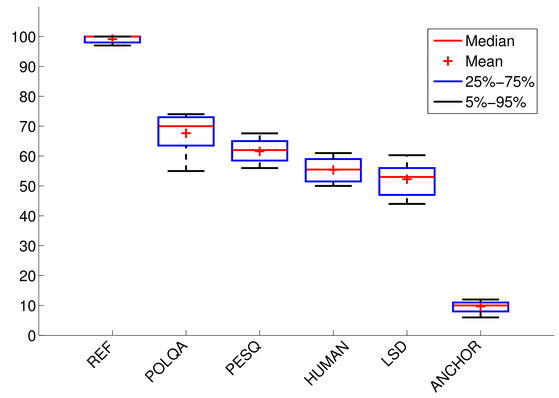
4 Results and Conclusions

- Significant difference in the resulting values were observed, 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 (5)).

	β	m_G	ξ_{H_1}	α_P	P_{TH}	α_{PSD}
$\mathbf{U}$	0.98	0.050	63.24	0.90	0.99	0.90
$\mathbf{L}$	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

Values resulting from the optimization methods for each parameter vector. The first two rows represent the bounds imposed on the parameter vector.

- Evaluation over the testing set provided $\Delta\text{MOS}\left(\hat{x}[k]|_{\mathbf{p}_{\text{POLQA}}}, \hat{x}[k]|_{\mathbf{p}_{\text{HUMAN}}}\right) \approx 0.2$ (as expected).
- In the MUSHRA test, $\mathbf{p}_{\text{POLQA}}$ and $\mathbf{p}_{\text{PESQ}}$ clearly achieves higher mean and median score of all the tuning methodologies.
- POLQA offers better performance compared to the one obtained through PESQ.
- Statistically significant difference between the combined scored obtained through $\mathbf{p}_{\text{POLQA}}$ and $\mathbf{p}_{\text{PESQ}}$ and the one obtained through $\mathbf{p}_{\text{LSD}}$ or $\mathbf{p}_{\text{HUMAN}}$.



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.

- Optimizing over perceptual objective criteria seems to work reasonably well in determining better solutions to the tuning problem.
- Future work:
  - extending the optimization framework to more complicated VQE algorithms including a larger number of blocks (e.g., linear and nonlinear echo canceler, comfort noise generators, multimic systems).
  - The larger the set of parameters that need tuning, the higher the expected gains in both quality and efficiency.

References

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

(2) *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 P.862, 2001.

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

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

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

(6) D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley, 1989.