IMPROVED KALMAN FILTERING FOR SPEECH ENHANCEMENT

Volodya Grancharov, Jonas Samuelsson and W. Bastiaan Kleijn

KTH (Royal Institute of Technology)
Department of Signals, Sensors and Systems
10044 Stockholm, Sweden
{Volodya.Grancharov, Jonas.Samuelsson, Bastiaan.Kleijn}@s3.kth.se

ABSTRACT

The Kalman recursion is a powerful technique for reconstruction of the speech signal observed in additive background noise. In contrast to Wiener filtering and spectral subtraction schemes, the Kalman algorithm can be easily implemented in both causal and noncausal form. After studying the perceptual differences between these two implementations we propose a novel algorithm that combines the low complexity and the robustness of the Kalman filter and the proper noise shaping of the Kalman smoother.

1. INTRODUCTION

The enhancement of noisy speech for mobile communications systems is a challenging and long-standing research problem. Kalman filtering is a general estimation technique applicable to the speech enhancement problem. The use of Kalman filtering for speech enhancement was first proposed in [1] and later extended to the more realistic colored noise case in [2]. A variety of Kalman filter implementations have been proposed for speech enhancement, some concerned with the speech model [3], some with parameter estimation schemes [4].

The Kalman algorithm in its causal form is the best linear signal estimator, in the mean-squared error sense, given the past and the present observations. However, for typical speech processing applications, future data is available and a noncausal implementation is possible. The optimal, in the mean-squared error sense, system that at each time point exploits all the available past and future data is the Kalman fixed-interval smoother. Removing the causality constraints naturally leads to a decrease in the mean-squared error. However, we do not focus on the error covariance, since it is known to be poorly correlated with human perception [5]. It is of practical interest to study the performance of the speech enhancement system in relation to the concept of "masking". Masking is the phenomenon that a weak signal is made inaudible in the vicinity of a strong signal. The valleys in a speech spectrum are the regions where in general the noise is not masked by the speech signal and, therefore stronger suppression is needed. We show that the causality constraint reduces the sharpness of the Kalman filter transfer function, and that the causality constraint introduces a degradation of the subjective quality. However, the causal implementation has low computational complexity and lacks the "musical" noise distortion that is typical for the noncausal algorithms. In this paper we developed a novel speech enhancement algorithm that inherits only the desired features of causal and noncausal Kalman algorithms.

This work was partially funded by Nokia Corporation.

2. KALMAN RECURSION

Let the speech signal recorded by the microphone be given by:

$$y_k = s_k + v_k, \tag{1}$$

where s_k is the sampled speech signal, and v_k an independent additive background noise. The Kalman algorithm provides a method to compute recursively the minimum mean-squared error estimate $\hat{s}_k$ from the available noisy observations.

2.1. The State-Space Model

In order to apply the Kalman filter, we model the speech and noise as autoregressive processes of model order p and q respectively:

$$s_k = \sum_{j=1}^{p} a_j s_{k-j} + w_k \tag{2}$$

$$v_k = \sum_{j=1}^{q} b_j v_{k-j} + u_k, \tag{3}$$

where w_k and u_k are white noise sequences. The speech and the noise model orders are typically set to ten for narrowband speech. The system of equations (1-3) can be represented in a state-space form:

$$\begin{aligned} \mathbf{x}_{k+1} &= \mathbf{F}\,\mathbf{x}_k + \mathbf{G}\,\mathbf{z}_k \\ y_k &= \mathbf{H}^T\,\mathbf{x}_k, \end{aligned} \tag{4}$$

$\mathbf{x}_k = [s_k\ s_{k-1}\ \ldots\ s_{k-p+1}\ v_k\ v_{k-1}\ \ldots\ v_{k-q+1}]^T$ is the $(p+q)$ dimensional state vector and $\mathbf{z}_k = [w_k\ u_k]^T$. The explicit expressions for $\mathbf{F}$, $\mathbf{G}$ and $\mathbf{H}$ are given below:

$$\mathbf{F} = \begin{pmatrix} \mathbf{F}_s & \mathbf{0}_{p,q} \\ \mathbf{0}_{p,q} & \mathbf{F}_v \end{pmatrix}$$

$$\mathbf{G} = \begin{pmatrix} 1\,0\,\cdots\,0 & 0\,0\,\cdots\,0 \\ \underbrace{0\,0\,\cdots\,0}_{p} & \underbrace{1\,0\,\cdots\,0}_{q} \end{pmatrix}^T$$

$$\mathbf{H} = \begin{pmatrix} \underbrace{1\,0\,\cdots\,0}_{p} & \underbrace{1\,0\,\cdots\,0}_{q} \end{pmatrix}^T,$$

and the speech transition matrix is given by:

$$\mathbf{F}_s = \begin{pmatrix} a_1 & a_2 & \cdots & a_{p-1} & a_p \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix}.$$

The noise transition matrix $\mathbf{F}_v$ is of the same form, except that its elements are the linear predictive coefficients b_j.

2.2. Causal Implementation

Using the state-space representation (4), the Kalman filter estimate becomes [2], [6]:

$$\begin{aligned} \hat{\mathbf{x}}_{k|k} &= \mathbf{F}\hat{\mathbf{x}}_{k-1|k-1} + \mathbf{K}_k e_k \\ e_k &= y_k - \mathbf{H}^T\mathbf{F}\hat{\mathbf{x}}_{k-1|k-1} \\ \mathbf{K}_k &= \mathbf{P}_{k|k-1}\mathbf{H}(\mathbf{H}^T\mathbf{P}_{k|k-1}\mathbf{H})^{-1} \\ \mathbf{P}_{k|k} &= [\mathbf{I} - \mathbf{K}_k\mathbf{H}^T]\mathbf{P}_{k|k-1} \\ \mathbf{P}_{k+1|k} &= \mathbf{F}\mathbf{P}_{k|k}\mathbf{F}^T + \mathbf{G}\mathbf{Q}\mathbf{G}^T \end{aligned} \tag{5}$$

where $\mathbf{K}_k$ is the Kalman gain and $\hat{\mathbf{x}}_{k|k}$ is the estimate of the state at a time index k, given observations up to a time index k. The prediction-error covariance matrix and the filtering-error covariance matrix are given by $\mathbf{P}_{k+1|k}$ and $\mathbf{P}_{k|k}$, and the noise covariance is given by $\mathbf{Q} = E\{\mathbf{z}_k\mathbf{z}_k^T\}$. The speech sample estimate can be obtained by $\hat{s}_k = [1\,0\,\cdots\,0]_{p+q}\hat{\mathbf{x}}_{k|k}$. The filter parameters $\mathbf{F}_s$, $\mathbf{F}_v$ and $\mathbf{Q}$ are updated on a frame-by-frame basis.

2.3. Noncausal Implementation

In Kalman filtering, the estimate $\hat{s}_k$ is causal and is based on the noisy measurement set $\{y_0, y_1, \ldots, y_k\}$. However, most communication systems permit a delay and the available noisy measurements are $\{y_0, y_1, \ldots, y_k, \ldots, y_{M-1}, y_M\}$. Here y_M is the last sample of the currently available speech frame. For off-line processing of speech, y_M can also be the last sample in the entire utterance.

The Kalman *fixed-interval smoother* is a well studied algorithm that can incorporate all the available future data. For the state-space model (4), for all l in the interval $0 \le l \le M$, we can determine the estimate $\hat{\mathbf{x}}_{l|M}$ given $\{y_0, y_1, \ldots, y_k, \ldots, y_{M-1}, y_M\}$ and the associated error covariance is $\mathbf{P}_{l|M} = E\{(\mathbf{x}_l - \hat{\mathbf{x}}_{l|M})(\mathbf{x}_l - \hat{\mathbf{x}}_{l|M})^T\}$. Using the innovations approach, the smoothed estimators can be obtained easily with the Bryson-Frazier recursion [6]. On a forward pass, over the interval $[0, M]$, we collect the quantities $\hat{\mathbf{x}}_{k|k}$, e_k, $\mathbf{P}_{k|k-1}$, $\mathbf{F}_{p,k} = \mathbf{F} - \mathbf{F}\mathbf{K}_k\mathbf{H}^T$ and $\mathbf{R}_{e,k} = \mathbf{H}^T\mathbf{P}_{k|k-1}\mathbf{H}$, according to the iteration (5). Then the adjoint variables $\lambda_{k|M}$ are computed via the backward recursion:

$$\lambda_{k|M} = \mathbf{F}_{p,k}^T\lambda_{k+1|M} + \mathbf{H}\mathbf{R}_{e,k}^{-1}e_k, \tag{6}$$

with an initial value $\lambda_{M+1|M} = 0$. Finally the smoothed state estimates are obtained as a weighted sum of filtered state estimates and adjoint variables:

$$\hat{\mathbf{x}}_{k|M} = \hat{\mathbf{x}}_{k|k} + \mathbf{P}_{k+1|k}\mathbf{F}_{p,k}^T\lambda_{k+1|M}. \tag{7}$$

The index M can be large, and the implementation of the Kalman fixed-interval smoother not feasible, because of high computational and storage demands. It is reasonable to study algorithms that incorporate only a "sufficient" amount of future data. The Kalman *fixed-lag smoother* determines for each time instant k and some fixed-lag $N \le M$, recursive equations for the state estimate $\hat{\mathbf{x}}_{k|k+N}$ given the observations $\{y_0, y_1, \ldots, y_k, \ldots, y_{k+N-1}, y_{k+N}\}$. The implementation of the fixed-lag smoother, used later in the simulation is also based on Bryson-Frazier formulas [6].

In the following sections, "delay" and "causality" will be related only to the type of Kalman algorithm implementation. Parameter estimation schemes, as well as pre- and post- processing algorithms are assumed to be completely separate from the filter implementation, and their delay is not discussed.

3. CAUSALITY AND PERCEIVED QUALITY

The differences between the discussed causal and noncausal implementations, in the mean-squared error sense, are well studied [6]. Unfortunately the mean-squared error is poorly correlated with the human perception of quality. Noise masking [7] is a well-known psychoacoustical property of the auditory system that has been applied successfully to speech coding [5] and enhancement [8]. Therefore it will be of interest to present the causal and noncausal algorithms in a form where we can easily see the differences from a noise-masking perspective.

3.1. Frequency Domain Representation

Let us assume a stationary signal, available at the infinite past and that the variables $\{\mathbf{R}_{e,k}, \mathbf{K}_k, \mathbf{F}_{p,k}\}$ have reached their steady-state values $\{\mathbf{R}_e, \mathbf{K}, \mathbf{F}_p\}$. After rearranging and taking the z-transform of the first equation in (5) we obtain:

$$\hat{x}_{KF}(z) = (\mathbf{I} - z^{-1}\mathbf{F})^{-1}\mathbf{K}\,e(z). \tag{8}$$

Thus, the transfer function from the innovations to the state estimate is:

$$H_c(z) = (\mathbf{I} - z^{-1}\mathbf{F})^{-1}\mathbf{K}. \tag{9}$$

Using equation (6) in a similar manner we find the transfer function from the innovations to the adjoint variable to be $(\mathbf{I} - z\mathbf{F}_p^T)^{-1}\mathbf{H}\mathbf{R}_e^{-1}$. It is then easy to see from equation (7) that the transfer function from innovations to the smoothed state estimate can be expressed as $H_c(z) + H_a(z)$, where:

$$H_a(z) = \mathbf{P}\mathbf{F}_p^T(\mathbf{I} - z\mathbf{F}_p^T)^{-1}\mathbf{H}\mathbf{R}_e^{-1}. \tag{10}$$

From the first two equations in (5), we see that $e(z) = L^{-1}(z)y(z)$, with the whitening filter:

$$L^{-1}(z) = \mathbf{I} - \mathbf{H}^T(\mathbf{I} - z^{-1}\mathbf{F}_p\mathbf{H}^T)^{-1}\mathbf{F}\mathbf{K}. \tag{11}$$

We finally obtain the transfer function of the Kalman filter and Kalman fixed-interval smoother in the desired form:

$$\begin{aligned} H_{KF}(z) &= H_c(z)L^{-1}(z) \\ H_{KS}(z) &= [H_c(z) + H_a(z)]L^{-1}(z). \end{aligned} \tag{12}$$

The Kalman smoother transfer function consists of a strictly causal term $H_c(z) = \sum_{j=1}^{\infty} z^{-j}\mathbf{F}^{j-1}\mathbf{K}$ and a strictly anticausal term $H_a(z) = \sum_{j=1}^{\infty} z^j\mathbf{P}(\mathbf{F}_p^T)^j\mathbf{H}\mathbf{R}_e^{-1}$. In the Kalman filter transfer function, the polynomial consists only of negative powers of z, since the anticausal term is truncated. As a result, the Kalman filter transfer function is not sufficiently sharp to model the speech spectral valleys, which are regions with lower contribution to the mean-squared error. The Kalman filter and the Kalman smoother match the spectral peaks essentially equally well and the difference is over the spectral valleys, where the Kalman filter leaves a significant amount of unmasked residual noise.

3.2. First-Order Model

To support the argumentation so far, we study a simple model with known analytical solution. We assume a stationary, infinitely long data segment and a first-order autoregressive model for the speech signal. Equation (2) then simplifies to:

$$s_k = as_{k-1} + w_k. \tag{13}$$

Suppose the signal is contaminated with stationary white Gaussian noise with known variance, according to equation (1), and we seek

the optimal mean-squared error linear estimate, in the causal and noncausal form.

Substituting the second equation into the first equation of the recursion (5) we obtain the general relation:

$$\hat{s}_{KF}(z) = (\mathbf{I} - z^{-1}(\mathbf{F} - \mathbf{K}\mathbf{H}^T\mathbf{F}))^{-1}\mathbf{K}\, y(z), \qquad (14)$$

and the transfer function of the steady-state Kalman filter for the given first-order model is:

$$H_{KF}(z) = \frac{K}{1 - a(1-K)z^{-1}}. \qquad (15)$$

Here the Kalman gain is:

$$K = \frac{\bar{P}}{\bar{P} + \sigma_v^2}, \qquad (16)$$

where $\bar{P}$ is the positive solution to the equation:

$$P^2 + [(1-a^2)\sigma_v^2 - \sigma_w^2]P - \sigma_v^2\sigma_w^2 = 0. \qquad (17)$$

The most straightforward manner to find a noncausal solution to the problem is by realizing that the transfer function of the optimal Kalman fixed-interval smoother, under the discussed above conditions, coincides with the well known noncausal Wiener filter solution:

$$H_{KS}(z) = \frac{P_s(z)}{P_y(z)} = \frac{\sigma_w^2}{\sigma_w^2 + \sigma_v^2(1-az)(1-az^{-1})}. \qquad (18)$$

Here $P_s(z)$ is the power spectrum of the clean speech signal and $P_y(z)$ is the power spectrum of the noisy observations. In Fig. 1

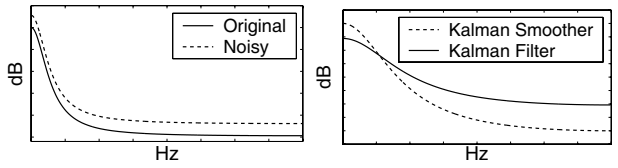

Fig. 1. Clean and noisy power spectrum for a first order signal and transfer functions of the Kalman fixed-interval smoother and the Kalman filter.

we illustrate the power spectrum of the clean and the noisy signals with the parameters $a = 0.8$, $\sigma_w^2 = 0.36$ and $\sigma_v^2 = 1.0$. In the same figure we present the frequency response of the transfer functions (18) and (15). Only in a case of an input signal with flat spectrum, i.e., $a = 0$, the causal and noncausal transfer functions are equivalent: $H_{KF}(z) = H_{KS}(z) = \frac{\sigma_w^2}{\sigma_w^2 + \sigma_v^2}$. In all other cases the Kalman fixed-interval smoother suppresses more noise than the Kalman filter for the high frequencies, where the signal energy is lower. By means of simulations we confirmed that this behavior generalizes to higher-order autoregressive models for speech.

4. IMPROVED KALMAN FILTERING BY PERCEPTUAL WEIGHTING

The conventional way to overcome the drawbacks of Kalman filtering is to remove the causality constraints and use a smoother instead. Even though theoretically optimal, this solution has some practical disadvantages, such as high computational complexity and "musical" noise distortion. Therefore we pose the problem of improving the perceptual performance of the Kalman filtering, while preserving its efficient causal structure. Since the Kalman filter is the optimal mean-squared error algorithm, further reduction of the error is not possible. However through proper weighting of the error covariance we can redistribute the error towards high energy speech regions, where it is less audible.

The idea of perceptual weighting of the error is extensively exploited in the speech coding [5]. The weighting filter that deemphasizes the formant structure of the speech signal is of the form:

$$H(z) = \frac{A(z/\gamma_1)}{A(z/\gamma_2)}, \qquad (19)$$

where $A(z)$ is the short term predictor filter and $1 \geq \gamma_1 \geq \gamma_2$.

The general scheme of the perceptual transformation can be seen in Fig. 2. The input to the Kalman filter is a signal with deemphasized formant structure and the inverse transform is applied to the processed signal. All the model parameters are estimated in a perceptual domain. The increased suppression of noise in the low

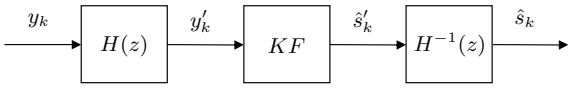

Fig. 2. Perceptual transformation for Kalman filtering by means of formant pre- and post- filter.

energy speech regions is clearly demonstrated in Fig. 3, where the autoregressive envelopes of original and processed signals for a representative speech frame are plotted. The proposed perceptual modification of the Kalman filter will not be optimal in the mean-squared error sense, but improvements in the perceived quality are confirmed in the next section by listening tests and objective measures that are well correlated with human perception.

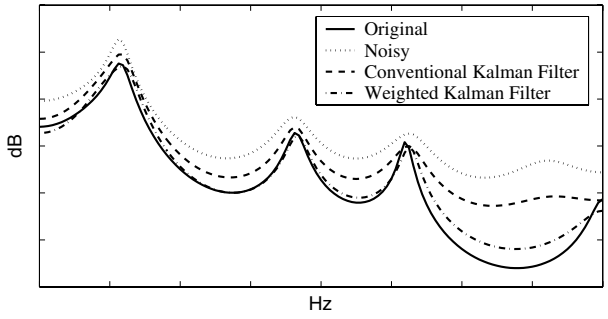

Fig. 3. Autoregressive envelopes of the standard and the proposed Kalman filter, with $\gamma_1 = 1.0$ and $\gamma_2 = 0.6$.

5. SIMULATIONS

In this section we perform simulations of the presented systems, with both "ideal" and estimated filter parameters. The test material consists of four clean speech sentences (two male and two female speakers), arbitrarily chosen from the TIMIT speech database [9]. The iterative scheme proposed in [2] was used to estimate the system parameters, for the cases where they were not known. The objective evaluation was performed in terms of the measures: SNR, spectral distortion (SD) [5], and PESQ [10]. The listening test setup was similar to the ITU recommendation ITU-R BS.1534 MUSHRA [11]. Ten listeners, not familiar with the systems, were asked to rate the systems on a scale from 0 to 100.

5.1. Optimal Delay for the Kalman Fixed-Lag Smoother

The noisy signal was processed with the Kalman filter, the Kalman fixed-interval smoother, and the fixed-lag smoother, with increasing lag. The required model parameters were estimated from the current frame of 20 ms, directly from the clean and noise signals.

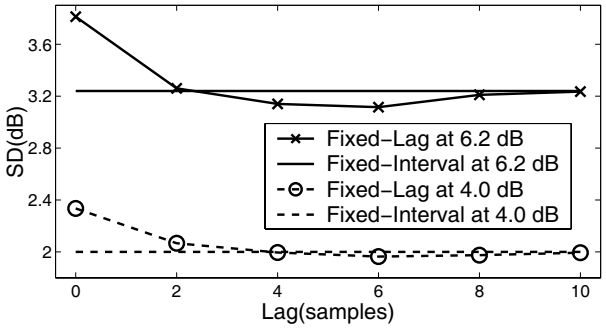

Fig. 4. Performance comparison in SD with white noise at input SD 6.2 dB and 4.0 dB.

The horizontal lines in Fig. 4 correspond to the optimal Kalman fixed-interval smoother with delay until the end of the speech signal. It is easy to see the significant difference between the performance of the Kalman filter with zero delay and the fixed-lag smoother. The simulations with SNR and PESQ exhibited similar behavior. From informal listening tests we conclude that perceptual equivalence between the fixed-interval smoother and the fixed-lag smoother is reached at lag $N = 10$, at 8 kHz sampling rate, despite of the faster convergence of the objective measures. A large difference between the causal and the delayed system is a strong indication that optimizing mean-squared error, under the causality constraints, results in an error distribution inconsistent with human perception.

5.2. Performance Comparison

The same test material and "ideal" conditions were used to evaluate the objective performance of the proposed weighted Kalman filter, Table 1. The weighting filter (19) was used with parameters $\gamma_1 = 1.0$ and $\gamma_2 = 0.6$. The fixed-lag smoother was used with lag $N = 10$. As expected, the SNR values are lower for the proposed implementation of the Kalman filter, since it minimizes mean-squared error in a transformed domain. In contrast, the SD and PESQ values, which are well correlated with human perception, were better for the proposed modification in Kalman filter.

	SNR	SD	PESQ
Kalman filter	14.8	2.6	2.924
Weighted Kalman filter	9.2	1.9	3.584
Fixed-lag smoother	16.8	2.2	3.750

Table 1. Performance comparison with ideal parameters [traffic noise at input SNR 10 dB].

The results from the listening tests, summarized in Table 2, show a clear advantage of the smoother over the causal implementations, for the case of known filter parameters. When the speech and noise model parameters were unknown, the weighted Kalman filter was ranked higher than the Kalman smoother and the conventional Kalman filter.

Noise Type	System type	Score
Traffic 5 dB	Kalman Filter	24.1
(Ideal)	Weighted Kalman Filter	25.4
	Fixed-Lag (N=10) Smoother	35.9
Traffic 5 dB	Kalman Filter	19.2
(Estimated)	Weighted Kalman Filter	22.5
	Fixed-Lag (N=10) Smoother	18.8

Table 2. Listeners' mean subjective scores for the proposed systems against the baseline Kalman filter and Kalman-fixed lag smoother. The 95% confidence interval does not exceed 2.4.

6. CONCLUSIONS

By means of theoretical analysis and simulations we showed that the Kalman smoother is more consistent with human perception than the Kalman filter, and found the minimum lag that guarantees the optimal performance for the Kalman fixed-lag smoother. Despite its disadvantages, the causal implementation has lower computational complexity and is more robust to errors in parameter estimation. The proposed weighted Kalman filter combines the proper noise shaping of the smoother and the robustness and low complexity of the causal implementation.

7. REFERENCES

[1] K. Paliwal and A. Basu, "A speech enhancement method based on Kalman filtering," *Proc. IEEE Int. Conf. Acous., Speech, Signal Processing*, vol. 12, pp. 177– 180, Apr. 1987.

[2] J. Gibson, B. Koo, and S. Gray, "Filtering of colored noise for speech enhancement and coding," *IEEE Trans. Signal Processing*, vol. 39, pp. 1732–1742, Aug. 1991.

[3] Z. Goh, K.-C. Tan, and B. Tan, "Kalman-filtering speech enhancement method based on a voiced-unvoiced speech model," *IEEE Trans. Speech, Audio Processing*, vol. 7, no. 5, pp. 510–524, Sept. 1999.

[4] S. Gannot, D. Burshtein, and E.Weinstein, "Iterative and sequential Kalman filter-based speech enhancement algorithms," *IEEE Trans. Speech, Audio Processing*, vol. 6, no. 4, pp. 373–385, July 1998.

[5] W. B. Kleijn and K. Paliwal, Eds., *Speech Coding and Synthesis*. Amsterdam: Elsevier Science Publishers, 1995.

[6] T. Kailath, A. Sayed, and B. Hassiby, *Linear Estimation*. New Jersey: Prentice Hall, 2000.

[7] B. C. J. Moore, *An Introduction to the Psychology of Hearing*. London: Academic Press, 1989.

[8] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Trans. Speech, Audio Processing*, vol. 7, pp. 126 – 137, March 1999.

[9] "DARPA-TIMIT," *Acoustic-phonetic continuous speech corpus, NIST Speech Disc 1-1.1*, 1990.

[10] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ) - A new method for speech quality assessment of telephone networks and codecs," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, pp. 749–752, 2001.

[11] "Method for the subjective assessment of intermediate quality level of coding systems, ITU-R Recommendation BS.1534," 2001.