SPEECH ENHANCEMENT USING A MODULATION DOMAIN KALMAN FILTER POST-PROCESSOR WITH A GAUSSIAN MIXTURE NOISE MODEL

Yu Wang and Mike Brookes

Department of Electrical and Electronic Engineering,
Exhibition Road, Imperial College London, UK
Email: {yw09, mike.brookes}@imperial.ac.uk

ABSTRACT

We propose a speech enhancement algorithm that applies a Kalman filter in the modulation domain to the output of a conventional enhancer operating in the time-frequency domain. We show that the prediction residual signal of the spectral amplitude errors at the output of the baseline MMSE enhancer do not follow a Gaussian distribution. Accordingly, the Kalman filter used in our enhancement algorithm combines a colored noise model with a Gaussian mixture model of the residual noise. We evaluate the performance of the speech enhancement algorithm on the core TIMIT test set and demonstrate that it gives consistent performance improvements over the baseline enhancer and over a previously proposed Kalman filter post-processor.

Index Terms— speech enhancement, post-processing, Kalman filter, Gaussian mixture model, modulation domain

1. INTRODUCTION

Over the past decades, many speech enhancement algorithms have been proposed in order to eliminate or reduce unwanted background noise. Enhancement algorithms in the time-frequency domain, such as [1] and [2], can be effective in reducing the noise and improving the signal-to-noise ratio (SNR) but they also distort the speech and introduce spurious artefacts known as musical noise. In order to solve this problem, several post-processing methods have been proposed that either filter the time-frequency gain function used within the enhancer or else act directly on its output signal. In [3], median filtering is applied to time-frequency cells that are identified as having a low probability of containing speech energy in order to eliminate the isolated peaks that characterise musical noise and in [4] musical noise in frames with low SNR is attenuated by smoothing the gain function of the baseline enhancer. Other techniques, such as cepstral smoothing [5] and Kalman filtering (KF) [6] have also been used to post-process the minimum mean square error (MMSE) spectral amplitude estimator [2] to reduce the musical noise and improve the quality of the enhanced speech.

The use of a KF was introduced in [7] for speech enhancement assuming the noise was white and in [8] this was later extended to colored noise. Recent interest in performing speech enhancement in the modulation domain [9, 10, 11] includes the algorithm described in [12] which applies the KF to the short-time modulation domain spectral amplitudes. The assumption made in the KF is that the prediction residual signals of both speech and noise are Gaussian distributed with zero-mean. However, we have found that the prediction residual signal of the spectral amplitude errors in the MMSE enhanced speech signal do not follow a Gaussian distribution. Therefore, extending the algorithm in [6] , we propose in this paper a KF post-processor in the modulation domain using a Gaussian mixture model (GMM) of the noise which is colored due to the overlap between frames. The rest of the paper is organized as follows: Sec. 2 gives the motivation and the probabilistic derivation of the GMM Kalman filter for colored noise and also describes the update procedure for the model parameters. In Secs 3 and 4, we evaluate the proposed algorithm and give our conclusions.

2. GMM KALMAN FILTER

2.1. Distribution of prediction error

In the conventional KF, the prediction residual signal of both speech and noise are assumed Gaussian distributed. However, after processing noisy speech by an MMSE enhancer, most of the stationary noise has been removed leaving behind some residual noise together with musical noise artefacts especially where the input noise power was high [13]. Because the musical noise is characterized by isolated spectral peaks in the spectrogram, it is difficult to predict in the modulation domain. As a result, the prediction errors associated with the musical noise may be very large, and the overall distribution of the prediction errors of the noise in the enhanced speech does not follow a Gaussian distribution. To illustrate this, we show in Fig. 1 the distribution of the normalized prediction error of the spectral amplitude errors in the MMSE enhanced speech in each time-frequency bin together with a fitted single Gaussian distribution (in red) and a 3-mixture GMM (in

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

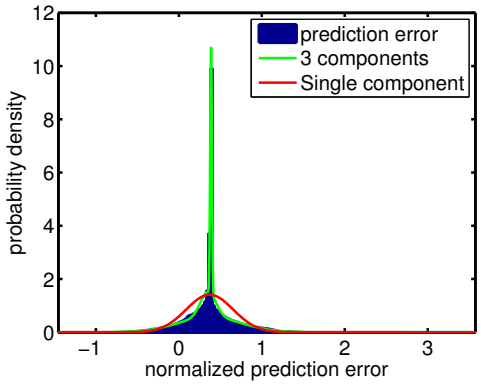

Fig. 1. Distribution of the normalized prediction error of the noise spectral amplitudes in MMSE-enhanced speech. The prediction errors are normalized by the RMS power of the noise predictor residual in the corresponding modulation frame.

green). The histogram shows the distribution over all time-frequency bins using the TIMIT core test set [14] corrupted by additive car noise at SNRs between –10 and +15 dB using the framing parameters from Sec. 3. The estimated noise amplitude trajectory in each frequency bin is represented by an autoregressive model whose prediction error is normalized by the root-mean-square (RMS) level of the noise predictor residual in the corresponding modulation frame. From the figure, we see that the overall prediction residual signal is not zero mean and does not follow a Gaussian distribution.

Based on the empirical prediction errors, we have extended the conventional colored noise KF to incorporate a GMM noise distribution. We use $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to denote a multivariate Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ and use $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ for its probability density at $\mathbf{x}$.

2.2. Derivation of GMM Kalman Filter

The diagram of the proposed algorithm is shown in Fig. 2. Following time-frequency domain enhancement, the spectral amplitude of the short-time Fourier transform (STFT) at time frame n and frequency bin k is given by $Y_{n,k} = X_{n,k} + W_{n,k}$ where $X_{n,k}$ is the amplitude of the clean speech signal and $W_{n,k}$ is the "noise" arising from a combination of acoustic noise and the enhancement artefacts. The output from the KF $\hat{X}_{n,k}$ is combined with the noisy phase spectrum $\theta_{n,k}$ and passed through an inverse-STFT (ISTFT) to create the output speech $\hat{x}(t)$. In this and the next subsection we will give the derivation of the GMM Kalman filter (GMMKF) and the parameter update procedure. Because each frequency bin, k, is processed independently and for clarity, we omit the frequency index below.

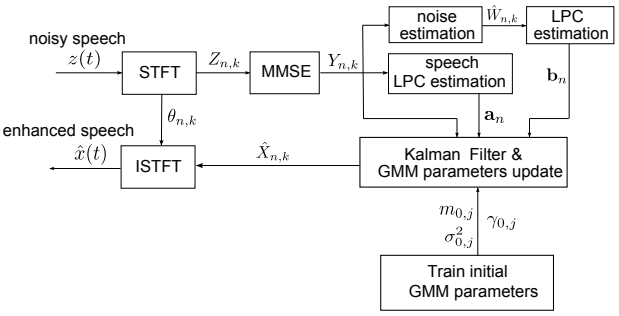

Fig. 2. Diagram of the proposed GMM KF algorithm

Our system model is

$$\mathbf{z}_{n+1} = \mathbf{A}_n \mathbf{z}_n + \mathbf{D}\mathbf{q}_n \tag{1}$$

$$y_{n+1} = \mathbf{c}^T \mathbf{z}_{n+1} \tag{2}$$

where $\mathbf{z}_n = [x_n \cdots x_{n-N+1}\, w_n \cdots w_{n-M+1}]^T$ is the $N+M$ dimensional state vector for both the speech, x_n, and noise, w_n, and $\mathbf{q}_n = [u_n\, v_n]^T$ contains the corresponding prediction residuals. The $(N+M) \times (N+M)$ transition matrix, $\mathbf{A}_n$, is in the form $\mathbf{A}_n = \begin{bmatrix} \mathbf{T}_a & \mathbf{0} \\ \mathbf{0} & \mathbf{T}_b \end{bmatrix}$ where $\mathbf{T}_a$ and $\mathbf{T}_b$ are the transition matrices for speech and noise respectively and $\mathbf{T}_a$ is given by $\mathbf{T}_a = \begin{bmatrix} -\mathbf{a}_n^T \\ \mathbf{I} \quad \mathbf{0} \end{bmatrix}$ where $\mathbf{a}_n$ is the vector of linear prediction (LPC) coefficients for the speech. $\mathbf{T}_b$ and $\mathbf{b}_n$ are the corresponding quantities for the noise. The $(N+M) \times 2$ matrix $\mathbf{D}$ is all zero except for $d_{1,1} = d_{N+1,2} = 1$. Likewise the column vector $\mathbf{c}$ is all zero except for $c_1 = c_{N+1} = 1$.

We represent the prediction residuals as a 2-element vector $\mathbf{q}_n$ with a Gaussian mixture distribution of J mixtures as

$$\mathbf{q}_n \sim \sum_{j=1}^{J} \gamma_n^{(j)} \mathcal{N}(\boldsymbol{\mu}_n^{(j)}, \boldsymbol{\Sigma}_n^{(j)}) \tag{3}$$

As in a conventional Kalman filter, we assume that the state vector at time n based on observations up to time n is Gaussian distributed $\mathbf{z}_n \sim \mathcal{N}(\mathbf{z}_{n|n}, \mathbf{P}_{n|n})$. Following the time update, the distribution of $\mathbf{z}_{n+1|n}$ becomes a Gaussian mixture $\sum_j \gamma_n^{(j)} \mathcal{N}(\mathbf{z}_{n+1|n}^{(j)}, \mathbf{P}_{n+1|n}^{(j)})$ where

$$\mathbf{z}_{n+1|n}^{(j)} = \mathbf{A}_n \mathbf{z}_{n|n} + \mathbf{D}\boldsymbol{\mu}_n^{(j)}$$

$$\mathbf{P}_{n+1|n}^{(j)} = \mathbf{A}_n \mathbf{P}_{n|n} \mathbf{A}_n^T + \mathbf{D}\boldsymbol{\Sigma}_n^{(j)} \mathbf{D}^T$$

Applying the constraint $\mathbf{c}^T \mathbf{z}_{n+1} = y_{n+1}$ changes the Gaussian mixture parameters as follows [15]

$$\mathbf{k}_{n+1}^{(j)} = \mathbf{P}_{n+1|n}^{(j)} \mathbf{c} (\mathbf{c}^T \mathbf{P}_{n+1|n}^{(j)} \mathbf{c})^{-1} \tag{4}$$

$$\mathbf{z}_{n+1|n+1}^{(j)} = \mathbf{z}_{n+1|n}^{(j)} + \mathbf{k}_{n+1}^{(j)} (y_{n+1} - \mathbf{c}^T \mathbf{z}_{n+1|n}^{(j)}) \tag{5}$$

$$\mathbf{P}_{n+1|n+1}^{(j)} = \mathbf{P}_{n+1|n}^{(j)} - \mathbf{k}_{n+1}^{(j)} \mathbf{c}^T \mathbf{P}_{n+1|n}^{(j)} \tag{6}$$

Finally, we collapse the GMM into a single Gaussian for the estimation of the state vector at time $n+1$

$$\xi_{n+1}^{(j)} = \frac{\gamma_n^{(j)} \mathcal{N}(y_{n+1}; \mathbf{c}^T \mathbf{z}_{n+1|n}^{(j)}, \mathbf{c}^T \mathbf{P}_{n+1|n}^{(j)} \mathbf{c})}{\sum_j \gamma_n^{(j)} \mathcal{N}(y_{n+1}; \mathbf{c}^T \mathbf{z}_{n+1|n}^{(j)}, \mathbf{c}^T \mathbf{P}_{n+1|n}^{(j)} \mathbf{c})} \quad (7)$$

$$\mathbf{z}_{n+1|n+1} = \sum_{j=1}^{J} \xi_{n+1}^{(j)} \mathbf{z}_{n+1|n+1}^{(j)} \quad (8)$$

$$\mathbf{P}_{n+1|n+1} = \sum_{j=1}^{J} \xi_{n+1}^{(j)} (\mathbf{P}_{n+1|n+1}^{(j)} + \mathbf{z}_{n+1|n+1}^{(j)} (\mathbf{z}_{n+1|n+1}^{(j)})^T) - \mathbf{z}_{n+1|n+1} \mathbf{z}_{n+1|n+1}^T \quad (9)$$

The quantity $\xi_{n+1}^{(j)}$ in (7) represents the posterior probability that $\mathbf{z}_{n+1}$ belongs to mixture j.

Thus we can use the new Kalman filter to process the residual noise in the MMSE enhanced speech because the GMM can be used to model the spectral amplitude errors in the enhanced speech.

2.3. Update of parameters

The spectral amplitudes, $Y_{n,k}$ are divided into overlapping modulation frames and autocorrelation LPC analysis [16] is performed in each modulation frame to obtain a vector of speech prediction coefficients, $\mathbf{a}$, and a residual power ρ_a^2. To obtain the corresponding noise coefficients, the sequence of spectral amplitudes, $Y_{n,k}$ is passed through a noise power spectrum estimator [17] before performing LPC analysis to obtain the noise predictor coefficients, $\mathbf{b}$, and the residual power ρ_b^2.

Within the noise GMM, (3), the speech residual component $u_n \sim \mathcal{N}(0, \rho_a^2)$ is identical in all mixture components but the normalized noise residual $e_n = v_n/\rho_b$ is modeled as a Gaussian mixture $e_n \sim \sum_j \gamma_n^{(j)} \mathcal{N}(m_n^{(j)}, \sigma_j^{2(j)})$. We model the normalized residual rather than the residual itself so that the GMM parameters are independent of the speech and noise amplitudes.

In order to update the GMM parameters we apply the noise predictor coefficients, $\mathbf{b}_n$, from the current modulation frame to the sequence of estimated noise spectral amplitudes to obtain a noise prediction error $e_n \rho_b$ for each acoustic frame n. The probability that e_{n+1} comes from model j is given by

$$p_{n+1}^{(j)} = \frac{\gamma_n^{(j)} \mathcal{N}(e_{n+1}; m_n^{(j)}, \sigma_n^{2(j)})}{\sum_j \gamma_n^{(j)} \mathcal{N}(e_{n+1}; m_n^{(j)}, \sigma_n^{2(j)})} \quad (10)$$

Because now the probability of the model given the observation error is known, we can update in each acoustic frame the effective number of observations $(O^{(j)})$, the sum of the observations $(S^{(j)})$ and the sum of the squared observations $(T^{(j)})$ as $O_{n+1}^{(j)} = p_{n+1}^{(j)} + \lambda O_n^{(j)}$, $S_{n+1}^{(j)} = p_{n+1}^{(j)} e_{n+1} + \lambda S_n^{(j)}$ and $T_{n+1}^{(j)} = p_{n+1}^{(j)} e_{n+1}^2 + \lambda T_n^{(j)}$, where λ is a forgetting factor. The parameters of each model can now be updated adaptively as [18]

$$m_{n+1}^{(j)} = S_{n+1}^{(j)} / O_{n+1}^{(j)} \quad (11)$$

$$\sigma_{n+1}^{2(j)} = T_{n+1}^{(j)} / O_{n+1}^{(j)} - m_{n+1}^{2(j)} \quad (12)$$

$$\gamma_{n+1}^{(j)} = \frac{O_{n+1}^{(j)}}{\sum_j O_{n+1}^{(j)}} = (1-\lambda) O_{n+1}^{(j)} \quad (13)$$

To initialize the model, we train a GMM with parameters $m_0^{(j)}$, $\sigma_0^{2(j)}$ and $\gamma_0^{(j)}$ offline on a large amount of data and set $O_0^{(j)} = m_0^{(j)}/(1-\lambda)$, $S_0^{(j)} = m_0^{(j)} O_0^{(j)}$ and $T_0^{(j)} = (\sigma_0^{2(j)} + m_0^{2(j)}) O_0^{(j)}$. To ensure stability of the update procedure, we impose lower bounds on $p^{(j)}$ and $\sigma^{2(j)}$ to prevent them becoming zero.

3. IMPLEMENTATION AND EVALUATION

3.1. Stimuli of experiments

In this section, we compare the performance of the proposed Kalman filter post-processor based on a GMM (KFGM) with the baseline MMSE enhancer from [2] and the modulation-domain Kalman filter post-processor (KFMD) from [6]. The initial GMM parameters are trained using a subset in the training set of the TIMIT database and using speech corrupted by white noise. The remaining algorithm parameters were chosen to optimize performance on a development subset of the TIMIT training database. The number of mixtures used is set as $J = 3$ and we select an acoustic frame length 16 ms with a 4 ms increment which gives a 250 Hz sampling frequency in the modulation domain. The speech and noise LPC models are determined from a modulation frame of 128 ms (32 acoustic frames) with a 16 ms frame increment and the model orders in the KFGM and KFMD algorithms for the speech and noise are $N = 3$ and $M = 4$ respectively. In the experiments, we use the core test set from the TIMIT database which contains 16 male and 8 female speakers each reading 8 distinct sentences (totalling 192 sentences) and the speech is corrupted by the F16 noise from the RSG-10 database [19] and street noise from the ITU-T test signals database [20] at $-10, -1, 0, 5, 10$ and 15 dB global SNR. A Hamming window is used in the STFT analysis and synthesis and the forgetting factor λ is set as $\lambda = 0.9$.

3.2. Performance evaluation

The performance of the algorithms is evaluated using both segmental SNR (segSNR) and the perceptual evaluation of speech quality (PESQ) measure defined in ITU-T P.862. All the measurement values are averaged over the 192 sentences in the TIMIT core test set. The average segSNR for the corrupted speech, baseline MMSE enhancer, the KFMD algorithm and the proposed KFGM algorithm is shown for F16

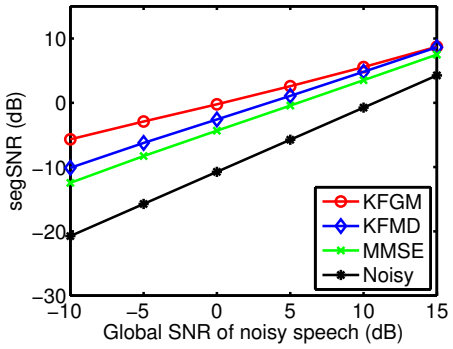

Fig. 3. Average segmental SNR of enhanced speech after processing by three algorithms versus the global SNR of the input speech corrupted by F16 aircraft noise (KFGM: proposed Kalman Filter post-processor with a Gaussian Mixture noise model; KFMD: modulation-domain Kalman filter post-processor from [6]; MMSE: MMSE enhancer from [2]).

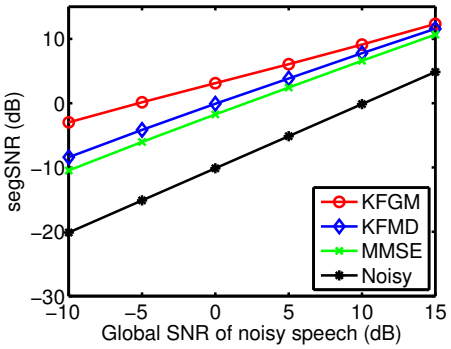

Fig. 4. Average segmental SNR of enhanced speech after processing by three algorithms versus the global SNR of the input speech corrupted by street noise.

noise in Fig. 3 as a function of the global SNR of the noisy speech. We see that at 15 dB global SNR all the algorithms give the same improvement in segSNR of about 3 dB. However, at 0 dB global SNR the proposed algorithm outperforms both reference algorithms by about 4 dB and 6 dB respectively. The equivalent graphs for street noise are shown in Fig. 4. We see that the overall trend in the results is the same and at 0 dB the proposed algorithm gives an additional improvement of 1.5 dB. The corresponding graphs for PESQ are shown in Fig. 5 for F16 noise and in Fig. 6 for street noise. In Figs 5 and 6, the average PESQ scores mirror the results seen for the segSNR. However, at high SNRs the proposed algorithm is also able to improve the PESQ, and we obtain an improvement of approximately 0.15 and 0.25 over the KFMD algorithm and MMSE enhancer respectively over a wide range of SNRs. In addition, informal listening tests also suggest that the proposed post-processing method is able to reduce the musical noise introduced by the MMSE enhancer.

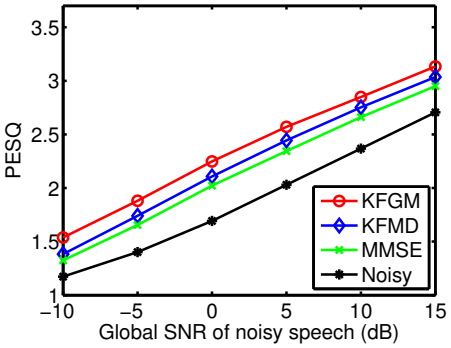

Fig. 5. Average PESQ quality of enhanced speech after processing by three algorithms versus the global SNR of the input speech corrupted by F16 aircraft noise.

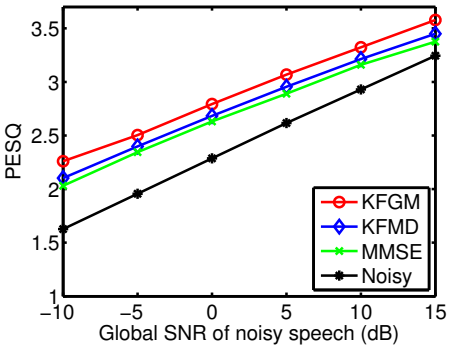

Fig. 6. Average PESQ quality of enhanced speech after processing by three algorithms versus the global SNR of the input speech corrupted by street noise.

4. CONCLUSION

In this paper we propose a new post-processor in the modulation domain using a GMM for modeling prediction error of the noise in the output of a conventional spectral amplitude MMSE enhancer. We have derived a KF that incorporates a GMM noise model and have also presented a method for adaptively updating the GMM parameters. We have evaluated our proposed post-processor using segSNR and PESQ and shown that the proposed method results in consistently improved performance when compared to both the baseline MMSE enhancer and a modulation-domain KF post-processor. The improvement in segmental SNR is over 4 dB at a global SNR of 0 dB while the PESQ score is increased by about 0.15 across a wide range of input global SNRs.

5. REFERENCES

[1] S. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoust., Speech, Signal Process.*, 27(2):113 – 120, April 1979.

[2] Y. Ephraim and D. Malah. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Trans. Acoust., Speech, Signal Process.*, 32(6):1109–1121, December 1984.

[3] Zenton Goh, Kah-Chye Tan, and T. G. Tan. Postprocessing method for suppressing musical noise generated by spectral subtraction. *IEEE Trans. Speech Audio Process.*, 6(3):287–292, May 1998.

[4] T. Esch and P. Vary. Efficient musical noise suppression for speech enhancement system. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4409–4412, April 2009.

[5] C. Breithaupt, T. Gerkmann, and R. Martin. Cepstral smoothing of spectral filter gains for speech enhancement without musical noise. *Signal Processing Letters, IEEE*, 14(12):1036–1039, December 2007.

[6] Y. Wang and M. Brookes. Speech enhancement using a robust Kalman filter post-processing in the modulation domain. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7457–7461, May 2013.

[7] K. Paliwal and A. Basu. A speech enhancement method based on Kalman filtering. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 177 – 180, April 1987.

[8] J. D. Gibson, B. Koo, and S. D. Gray. Filtering of colored noise for speech enhancement and coding. *IEEE Trans. Signal Process.*, 39(8):1732–1742, 1991.

[9] K. Paliwal, K. Wojcicki, and B. Schwerin. Single-channel speech enhancement using spectral subtraction in the short-time modulation domain. *Speech Communication*, 52(5):450–475, 2010.

[10] T. H. Falk, S. Stadler, W. B. Kleijn, and W. Y. Chan. Noise suppression based on extending a speech-dominated modulation band. In *Proc. Interspeech Conf.*, pages 970–973, August 2007.

[11] J. G. Lyons and K. K. Paliwal. Effect of compressing the dynamic range of the power spectrum in modulation filtering based speech enhancement. In *Proc. Interspeech Conf.*, pages 387–390, September 2008.

[12] S. So and K. K. Paliwal. Modulation-domain Kalman filtering for single-channel speech enhancement. *Speech Communication*, 53(6):818–829, July 2011.

[13] O. Cappe. Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor. *IEEE Trans. Speech Audio Process.*, 2(2):345–349, April 1994.

[14] John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett, Nancy L. Dahlgren, and Victor Zue. TIMIT acoustic-phonetic continuous speech corpus. Corpus LDC93S1, Linguistic Data Consortium, Philadelphia, 1993.

[15] Mike Brookes. The matrix reference manual. http://www.ee.imperial.ac.uk/hp/staff/dmb/matrix/intro.html, 1998-2013.

[16] J. Makhoul. Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4):561 – 580, April 1975.

[17] T. Gerkmann and R. C. Hendriks. Unbiased MMSE-based noise power estimation with low complexity and low tracking delay. *IEEE Trans. Audio, Speech, Lang. Process.*, 20(4):1383 –1393, May 2012.

[18] D. A. Reynolds, T. F. Quatieri, and R. Dunn. Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 10(1–3):19–41, 2000.

[19] H. J. M. Steeneken and F. W. M. Geurtsen. Description of the RSG.10 noise data-base. Technical Report IZF 1988–3, TNO Institute for perception, 1988.

[20] ITU-T P.501. Test signals for use in telephonometry, August 1996.