EUSIPCO 2013 1569743759

A SUBSPACE METHOD FOR SPEECH ENHANCEMENT IN THE MODULATION DOMAIN

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 present a modulation-domain speech enhancement algorithm based on a subspace method. We demonstrate that in the modulation domain, the covariance matrix of clean speech is rank deficient. We also derive a closed-form expression for the modulation-domain covariance matrix of colored noise in each frequency bin that depends on the analysis window shape and the noise power spectral density. Using this, we combine a noise power spectral density estimator with an efficient subspace method using a time domain constrained (TDC) estimator of the clean speech spectral envelope. The performance of the novel enhancement algorithm is evaluated using the PESQ measure and shown to outperform competitive algorithms for colored noise.

Index Terms- speech enhancement, subspace, modulation domain, covariance matrix estimation

1. INTRODUCTION

With the increasing use of hands-free telephony, especially within cars, it is often the case that speech signals are contaminated by the addition of unwanted background acoustic noise. The goal of a speech enhancement algorithm is to reduce or eliminate this background noise without distorting the speech signal. Over the past several decades, numerous speech enhancement algorithms have been proposed including a class of algorithms, introduced in [1], in which the space of noisy speech vectors is decomposed into a *signal subspace* containing both speech and noise and a *noise subspace* containing only noise. The clean speech is estimated by projecting the noisy speech vectors onto the signal subspace using a linear estimator that minimizes the speech signal distortion while applying either a time domain constraint (TDC) or spectral domain constraint (SDC) to the residual noise energy. The enhancer in [1], which assumed white or whitened noise, was extended to cope with colored noise in [2]. Different decompositions were applied in [3] to speech-dominated and noise-dominated frames since the latter do not require prewhitening. In a generalization of the approach, [4] apply a non-unitary transformation to the noisy speech vectors that simultaneously diagonalizes the covariance matrices of both speech and colored noise.

There is increasing evidence that information in speech is carried by the modulation of the spectral envelopes rather than by the envelopes themselves [5, 6, 7]. Consequently several recently proposed enhancers act in the short-time modulation domain using minimum mean-square error (MMSE) estimation [8], spectral subtraction [9] or Kalman filtering [10, 11].

This paper extends the subspace enhancement approach to the modulation domain and shows that, in this domain, the normalized noise covariance matrix can be taken to be fixed. The remainder of this paper is organized as follows. In Sec. 2 the principle of enhancement in the short-time modulation domain is described and in Sec. 3 we derive the noise covariance matrix estimate in this domain. Finally in Sec. 4 and Sec. 5 we evaluate the algorithm and give our conclusions.

2. SUBSPACE METHOD IN THE SHORT-TIME MODULATION DOMAIN

The block diagram of the proposed modulation-domain subspace enhancer is shown in Fig. 2. The noisy speech $y(r)$ is first transformed into the acoustic domain using a short-time Fourier transform (STFT) to obtain a sequence of spectral envelopes $Y(n,k)e^{j\theta(n,k)}$ where $Y(n,k)$ is the spectral amplitude of frequency bin k in frame n. The sequence $Y(n,k)$ is now divided into overlapping windowed modulation frames of length L with a frame increment J giving $Y_l(n,k) = p(n)Y(lJ+n,k)$ for $n = 0, \cdots, L-1$ where $p(n)$ is a Hamming window. A TDC subspace enhancer is applied independently to each frequency bin within each modulation frame to obtain the estimated clean speech spectral amplitudes $\widehat{S}_l(n,k)$ in frame l. The modulation frames are combined using overlap-addition to obtain the estimated clean speech envelope sequence $\widehat{S}(n,k)$ and these are then combined with the noisy speech phases $\theta(n,k)$ and an inverse STFT (ISTFT) applied to give the estimated clean speech signal $\hat{s}(r)$.

Following [12, 10] we assume a linear model in the spectral amplitude domain

$$Y_l(n,k) = S_l(n,k) + W_l(n,k) \tag{1}$$

where S and W denote the spectral amplitudes of clean speech and noise respectively. Since each frequency bin is

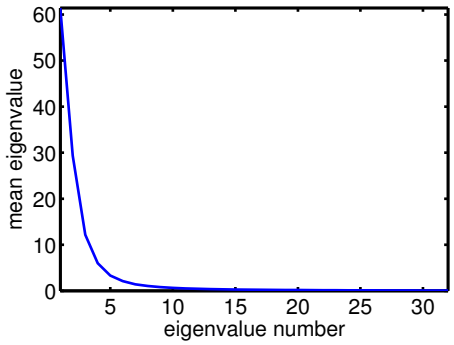

Fig. 1. Mean eigenvalues of covariance matrix of clean speech.

processed independently, we will omit the frequency index, k, in the remainder of this section. We define the noisy speech vector $\mathbf{y}_l = \begin{bmatrix} Y_l(0) & \cdots & Y_l(L-1) \end{bmatrix}^T$ and similarly for $\mathbf{s}_l$ and $\mathbf{w}_l$. The key assumption underlying the subspace enhancement method is that the covariance matrix of the clean speech vector, $\mathbf{s}_l$, is rank-deficient. To illustrate the validity of this, we show in Fig. 1 the ordered eigenvalues of the modulation domain speech vector covariance matrix of speech vector, $\mathbf{R}_S = \left\langle \mathbf{s}_l \mathbf{s}_l^T \right\rangle$, averaged over the TIMIT core test set using the framing parameters defined in Sec. 4.1 with a modulation frame length $L = 32$, where $\langle \ldots \rangle$ denotes the expected value. We see that the eigenvalues decrease rapidly and that 97% of the speech energy is included in the first 10 eigenvalues. We note that this low-rank assumption is also implicit in the use of a low-order LPC model in the modulation domain in [13, 11].

If $\mathbf{R}_Y$ and $\mathbf{R}_W$ are defined similarly to $\mathbf{R}_S$, we can, if we know $\mathbf{R}_W$, perform the eigen-decomposition

$$\mathbf{R}_W^{-\frac{1}{2}} \mathbf{R}_Y \mathbf{R}_W^{-\frac{1}{2}} = \mathbf{R}_W^{-\frac{1}{2}} \mathbf{R}_S \mathbf{R}_W^{-\frac{1}{2}} + \mathbf{I} = \mathbf{U}\mathbf{D}\mathbf{U}^T \qquad (2)$$

where $\mathbf{R}_W^{\frac{1}{2}}$ is the positive definite square root of $\mathbf{R}_W$. From this we can estimate the whitened clean speech eigenvalues as

$$\Lambda = \max(\mathbf{D} - \mathbf{I}, 0) \qquad (3)$$

We will estimate the clean speech vector from the noisy vector using a linear estimator, $\mathbf{H}_l$, as

$$\hat{\mathbf{s}}_l = \mathbf{H}_l \mathbf{y}_l \qquad (4)$$

It is shown in [2] that the optimal TDC linear estimator is given by

$$\mathbf{H}_l = \mathbf{R}_W^{\frac{1}{2}} \mathbf{U} \Lambda (\Lambda + \mu \mathbf{I})^{-1} \mathbf{U}^T \mathbf{R}_W^{-\frac{1}{2}} \qquad (5)$$

where μ controls the tradeoff between speech distortion and noise suppression.

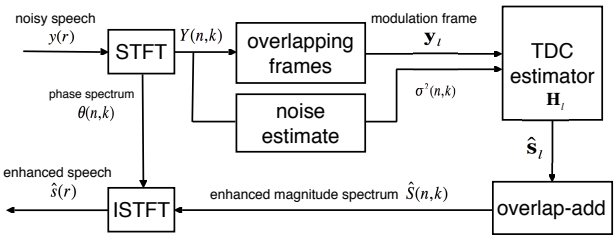

Fig. 2. Diagram of proposed short-time modulation domain subspace enhancer.

We can interpret the action of the estimator in (5) as first whitening the noise with $\mathbf{R}_W^{-\frac{1}{2}}$ and then applying a Karhunen-Loève transform (KLT), $\mathbf{U}^T$ to perform the subspace decomposition. In the transform domain, the gain matrix, $\Lambda(\Lambda + \mu\mathbf{I})^{-1}$, projects the vector into the signal subspace and attenuates the noise by a factor controlled by μ, discussed in Sec. 4.1. A detailed derivation of (5) is given in [1] and [2].

3. NOISE COVARIANCE MATRIX ESTIMATION

We now consider the estimation of the noise covariance matrix $\mathbf{R}_W$. For quasi-stationary noise, $\mathbf{R}_W$ will be a symmetric Toeplitz matrix whose first column is given by the autocorrelation vector $\mathbf{a}(k) = \begin{bmatrix} a(0,k) & \cdots & a(L-1,k) \end{bmatrix}^T$ where $a(\tau, k) = \langle W(n,k) W(n+\tau, k) \rangle$. We begin by determining $a(\tau, k)$ for the case when $w(r)$ is white noise and then extend this to colored noise.

First suppose $w(r) \sim N(0, \nu^2)$ is a zero-mean Gaussian white noise signal. If the acoustic frame length is R samples with a frame increment of M samples, the output of the initial STFT stage in Fig. 2 is

$$\widetilde{W}(n,k) = \sum_{r=0}^{R-1} w(nM + r) q(r) e^{-2\pi j \frac{rk}{R}} \qquad (6)$$

where $q(r)$ is the window function and the complex spectral coefficients, $\widetilde{W}(n,k)$, have a zero-mean complex Gaussian distribution [14]. The expectation $\left\langle \widetilde{W}(n,k) \widetilde{W}(n+\tau,k)^* \right\rangle$, where $*$ denotes complex conjugation, is given by

$$\begin{aligned} &\left\langle \widetilde{W}(n,k) \widetilde{W}(n+\tau,k)^* \right\rangle \\ &= \left\langle \sum_{r,s=0}^{R-1} w(nM+r) q(r) w(nM+s+\tau M) q(s) e^{-2\pi j \frac{(r-s)k}{R}} \right\rangle \\ &= \nu^2 \sum_{r=0}^{R-1} q(r) q(r - \tau M) e^{-2\pi j \frac{\tau M k}{R}} \end{aligned} \qquad (7)$$

since, for white noise,

$$\langle w(nM+r) w(nM+s+\tau M) \rangle = \nu^2 \delta(r - s - \tau M).$$

By setting $\tau = 0$, we can therefore obtain the spectral power in any frequency bin as

$$\sigma^2 = \left\langle \left| \widetilde{W}(n,k) \right|^2 \right\rangle = \nu^2 \sum_{r=0}^{R-1} q^2(r) \tag{8}$$

Defining

$$\rho(\tau,k) = \frac{\sum_{r=0}^{R-1} q(r)q(r-\tau M)e^{-2\pi j \frac{\tau M k}{R}}}{\sum_{r=0}^{R-1} q^2(r)}$$

we can now use (7) and (8) to write

$$\left\langle \widetilde{W}(n,k)\widetilde{W}(n+\tau,k)^* \right\rangle = \sigma^2 \rho(\tau,k)$$

where $\rho(\tau,k)$ depends on the window, $q(r)$, but not on the noise variance ν^2.

We now have obtained the autocorrelation sequence of the short-time Fourier coefficients $\left\langle \widetilde{W}(n,k)\widetilde{W}(n+\tau,k)^* \right\rangle$, from [15, pp. 95-97] we can further obtain the autocorrelation sequence of their magnitudes as

$$\begin{aligned} a(\tau,k) &= \langle W(n,k)W(n+\tau,k) \rangle \\ &= \left\langle \left| \widetilde{W}(n,k) \right| \left| \widetilde{W}(n+\tau,k) \right| \right\rangle \\ &= \frac{\pi}{4}\sigma^2 \times {}_2F_1(-\frac{1}{2},-\frac{1}{2},1;|\rho(\tau,k)|^2) \end{aligned} \tag{9}$$

where ${}_2F_1(\cdots)$ is the hypergeometric function [16] defined by

$${}_2F_1(m,n,o;z) = \sum_{k=0}^{\infty} \frac{(m)_k (n)_k}{(o)_k} \frac{z^k}{k!} \tag{10}$$

where $(m)_k = \frac{1}{m+k} \prod_{r=1}^{k+1} (m+r-1)$ is the rising Pochhammer symbol.

Therefore, if we define

$$\mathbf{a}_0(k) = \sigma^{-2} \left[\begin{array}{ccc} a(0,k) & \cdots & a(L-1,k) \end{array} \right]^T$$

and $\mathbf{R}_0(k)$ is a symmetric Toeplitz matrix with $\mathbf{a}_0(k)$ is the first column, we can write

$$\mathbf{R}_W(k) = \sigma^2 \mathbf{R}_0(k) \tag{11}$$

where $\mathbf{R}_0(k)$ does not depend on σ^2.

If we now assume that $w(r)$ is quasi-stationary colored noise with a correlation time that is small compared with the acoustic frame length, $\widetilde{W}(n+\tau,k)$ will be multiplied by a factor that depends on k but not on τ [17]. In this case, the previous analysis still applies but, for frame l, (11) now becomes

$$\mathbf{R}_W(k) = \sigma_l^2(k) \mathbf{R}_0(k) \tag{12}$$

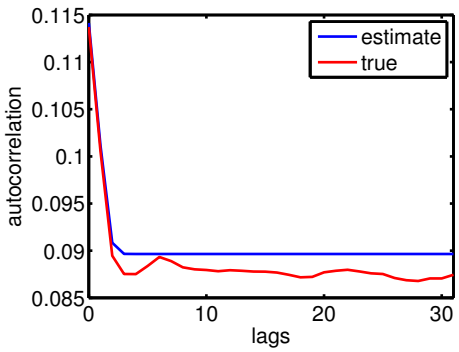

Fig. 3. Estimated and true value of the average autocorrelation sequence in one modulation frame.

where $\sigma_l^2(k) = \left\langle W^2(lJ,k) \right\rangle$ is the noise periodogram and, as shown above, $\mathbf{R}_0(k)$ is independent of the noise power spectrum. This means that we are able to estimate $\mathbf{R}_W(k)$ directly from an estimate of $\sigma_l^2(k)$ which can be obtained from the noisy speech signal, $y(r)$, using a noise power spectrum estimator such as [18] or [19].

Substituting (12) into (2)-(5), we obtain

$$\begin{aligned} \mathbf{R}_0^{-\frac{1}{2}} \mathbf{R}_Y \mathbf{R}_0^{-\frac{1}{2}} &= \mathbf{U}\overline{\mathbf{D}}\mathbf{U}^T \\ \overline{\Lambda} &= \max(\overline{\mathbf{D}} - \sigma_l^2(k)\mathbf{I}, 0) \\ \mathbf{H}_l &= \mathbf{R}_0^{\frac{1}{2}} \mathbf{U}\overline{\Lambda}(\overline{\Lambda} + \mu\sigma_l^2(k)\mathbf{I})^{-1}\mathbf{U}^T\mathbf{R}_0^{-\frac{1}{2}} \end{aligned}$$

in which the whitening transformation, $\mathbf{R}_0^{-\frac{1}{2}}$, can be precomputed since it depends only on the window, $h(r)$, and is independent of the noise power spectrum. In addition, because the matrix $(\overline{\Lambda} + \mu\sigma_l^2(k)\mathbf{I})$ is a diagonal matrix whose inverse is straightforward to calculate, the computational complexity of the estimator is greatly reduced.

To confirm the validity of the analysis we have evaluated the autocorrelation vector, $\mathbf{a}$, for the 'f16' noise in the RSG-10 database [20] using the framing parameters given in Sec. 4.1 with a modulation frame length $L = 32$. Figure 3 shows the true autocorrelation averaged over all k together with the autocorrelation from (9) using the true noise periodogram. We see that the two curves match very closely and that for $\tau \geq \frac{R}{J} = 4$, the STFT analysis windows do not overlap and so $a(\tau,k)$ is constant.

4. EXPERIMENTAL RESULTS

4.1. Implementation and Stimuli

In this section, we compare our proposed modulation domain subspace (MDSS) enhancer with the TDC version of the time-domain subspace (TDSS) enhancer[1] from [4] and the

[1]The Matlab implementation can be found in [21]

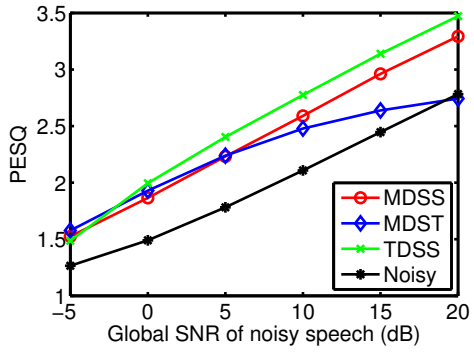

Fig. 4. Average PESQ values comparing different algorithms, where speech signals are corrupted by white noise at different SNR levels.

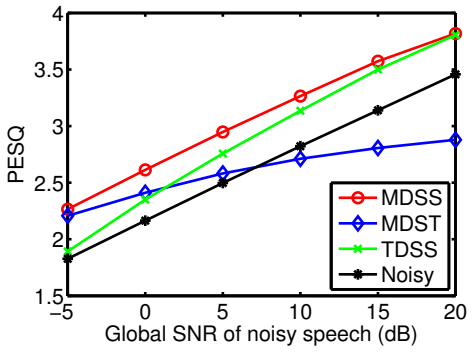

Fig. 5. Average PESQ values comparing different algorithms, where speech signals are corrupted by factory noise at different SNR levels.

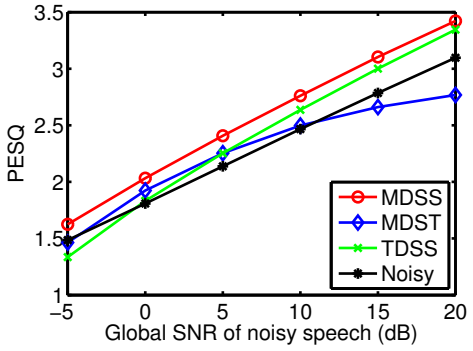

Fig. 6. Average PESQ values comparing different algorithms, where speech signals are corrupted by babble noise at different SNR levels.

modulation-domain spectral subtraction (MDST) enhancer[2] from [9] using the default parameters. In our experiments, we used the core test set from the TIMIT database [22] which contains 16 male and 8 female speakers each reading 8 distinct sentences (totalling 192 sentences) corrupted by 'white', 'factory2' and 'babble' noise from [20] at -5, 0, 5, 10, 15 and 20 dB signal-to-noise ratio (SNR). The algorithm parameters were determined by optimizing performance on a subset of the TIMIT training set. All speech and noise signals were downsampled to 8 kHz. The estimator in (5) was used to process each modulation frame of length 128ms with 16ms increment and the acoustic frames are 16ms long with 4ms increment ($L = 32$, $J = 4$, $R = 128$, $M = 32$). A Hamming window is applied for analysis and synthesis in both acoustic domain and modulation domain. Additionally, the noise power spectrum was estimated using the algorithm in [19, 23] and, following [4], the factor μ in (5) was selected as

$$\mu = \begin{cases} 5 & SNR_{dB} \leq -5 \\ \mu_0 - (SNR_{dB})/s & -5 < SNR_{dB} < 20 \\ 1 & SNR_{dB} \geq 20 \end{cases}$$

where $\mu_0 = 4.2$, $s = 6.25$, $SNR_{dB} = 10log_{10}(tr(\Lambda)/L)$.

To avoid any of the estimated spectral amplitudes in $\hat{\mathbf{s}}_l$ becoming negative, we set a floor equal to 20 dB below the corresponding noisy spectral amplitudes in $\mathbf{y}_l$ so that (4) now becomes

$$\hat{\mathbf{s}}_l = \max(\mathbf{H}_l \mathbf{y}_l, 0.1\mathbf{y}_l) \tag{13}$$

4.2. Experimental results

The performance of the three speech enhancers are evaluated and compared using the perceptual evaluation of speech quality (PESQ) measure defined in ITU-T P.862, averaged over the 192 sentences in the core TIMIT test set. The experimental results are shown in Fig. 4 to Fig. 6, for noisy speech corrupted by white noise, factory noise and babble noise respectively at different global SNRs, and the corresponding enhanced speech by the three enhancers mentioned above. We can see that, for colored noise, the proposed MDSS enhancer performs better than the other two enhancers, especially at low SNRs which gives a PESQ improvement of more than 0.2 over a wide range of SNRs. For white noise, the TDSS enhancer is better than the MDSS enhancer except at very low SNRs.

[2]The Matlab software is available online at url: http://maxwell.me.gu.edu.au/spl/research/modspecsub/

5. CONCLUSIONS

In this paper we have presented a speech enhancement algorithm using a subspace decomposition technique in the short-time modulation domain. We have derived a closed-form expression for the modulation-domain covariance matrix of quasi-stationary colored noise that depends on the STFT analysis window and the noise power spectral density. We have evaluated the performance of our proposed enhancer

using PESQ and shown that, for colored noise, it outperforms a time-domain subspace enhancer and modulation-domain spectral-subtraction enhancer.

6. REFERENCES

[1] Y. Ephraim and H. L. Van Trees. A signal subspace approach for speech enhancement. *IEEE Trans. Speech Audio Process.*, 3(4):251–266, July 1995.

[2] H. Lev-Ari and Y. Ephraim. Extension of the signal subspace speech enhancement approach to colored noise. *IEEE Signal Process. Lett.*, 10(4):104–106, April 2003.

[3] U. Mittal and N. Phamdo. Signal/noise KLT based approach for enhancing speech degraded by colored noise. *IEEE Trans. Speech Audio Process.*, 8(2):159–167, March 2000.

[4] Y. Hu and P. C. Loizou. A generalized subspace approach for enhancing speech corrupted by colored noise. *IEEE Trans. Speech Audio Process.*, 11(4):334–341, July 2003.

[5] H. Hermansky. The modulation spectrum in the automatic recognition of speech. In *Automatic Speech Recognition and Understanding, Proceedings.*, pages 140 –147, December 1997.

[6] R. Drullman, J. M. Festen, and R. Plomp. Effect of reducing slow temporal modulations on speech reception. *J. Acoust. Soc. Am.*, 95(5):2670–2680, 1994.

[7] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen. A short-time objecitve intelligibility measure for time-frequency weighted noisy speech. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4214–4217, 2010.

[8] Kuldip Paliwal, Belinda Schwerin, and Kamil Wójcicki. Speech enhancement using a minimum mean-square error short-time spectral modulation magnitude estimator. *Speech Commun.*, 54:282–305, February 2012.

[9] Kuldip Paliwal, Kamil Wójcicki, and Belinda Schwerin. Single-channel speech enhancement using spectral subtraction in the short-time modulation domain. *Speech Commun.*, 52:450–475, May 2010.

[10] S. So and K.K. Paliwal. Suppressing the influence of additive noise on the kalman gain for low residual noise speech enhancement. *Speech Communication*, 53(3):355–378, 2011.

[11] Yu Wang and Mike Brookes. Speech enhancement using a robust Kalman filter post-processor in the modulation domain. to appear in the Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP), May 2013.

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

[13] S. So, K.K. Wójcicki, and K.K. Paliwal. Single-channel speech enhancement using Kalman filtering in the modulation domain. In *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[14] 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.

[15] Kenneth S Miller. *Complex stochastic processes: an introduction to theory and application*. Addison-Wesley Publishing Company, Advanced Book Program, 1974.

[16] F. Olver, D. Lozier, R. F. Boiszert, and C. W. Clark, editors. *NIST Handbook of Mathematical Functions: Companion to the Digital Library of Mathematical Functions*. Cambridge University Press, 2010.

[17] Y. Avargel and I. Cohen. On multiplicative transfer function approximation in the short-time Fourier transform domain. *IEEE Signal Process. Lett.*, 14(5):337–340, 2007.

[18] R. Martin. Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Trans. Speech Audio Process.*, 9:504–512, July 2001.

[19] 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.

[20] 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.

[21] P. C. Loizou. Speech databases and MATLAB codec. In *Speech Enhancement Theory and Practice*, chapter Appendix C, pages 589–599. Taylor & Francis, 2007.

[22] J. S. Garofolo. Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database. Technical report, National Institute of Standards and Technology (NIST), Gaithersburg, Maryland, December 1988.

[23] D. M. Brookes. VOICEBOX: A speech processing toolbox for MATLAB. http://www.ee.imperial.ac.uk/hp/staff/dmb/voicebox/voicebox.html, 1998-2012.