Int. Conf. on Signal, Image Processing Communication & Automation, ICSIPCA

Blind Dereverberation for Speech Signal using Multichannel, A Review

Praveen P B[1], Dr Aravinda H S[2]
[1]Research scholar, JSS Academy, Bengaluru,
praveen.patelb@gmail.com
[2]Professor, HOD, Department of E & C, JSS Academy, Bengaluru
aravindhs1@gmail.com

***Abstract*—Quality of speech captured in room degrades severely due to reverberation. This in turn degrades the performance of the applications like, automatic speech recognition, telecommunication etc. Removing/Minimizing this reverberation effect is called dereverberation. In most of the cases, Room impulse response (RIR) is unknown, de-reverberation in such case is called blind de-reverberation. Several blind dereverberation techniques have been proposed. In this paper we are reviewing three papers. For each paper we have discussed its algorithm, complexity and limitation. In first review [1], we discuss mathematical model and algorithm for speech enhancement using LP residual cepstrum. In second review [2], we discussed speech enhancement using statistical model of late reverberation and in last review [3] we discussed speech enhancement using linear prediction and prediction filter.**

***Index Terms*— Blind dereverberation, multichannel speech enhancement, distant speech enhancement.**

I. INTRODUCTION

A speech signal captured using distant microphone degrades the speech quality severely due to reverberation. This degradation of speech affect severely the performance of speech application like automatic speech recognition, speaker detection, hands free telephony, hearing aids etc. De-reverberation using de-convolution with room impulse response (RIR) is the general idea behind improving the speech quality. In most of the cases, Room impulse response (RIR) is unknown, de-reverberation in such case is called blind de-reverberation. There are several technique proposed for blind reverberation using single and multiple microphone. In this review paper we are going review the 3 papers [1] [2] [3], we will describe the algorithm proposed, complexity and limitation.

If s(t) is the clean speech which is captured by a microphone located at distance "d" meter in a large room of impulse response H(t), then u(t), the output of microphone can be described mathematically as

$$u(t) = s(t) * H(t) \quad (1)$$

In frequency domain

$$u(e^{jw}) = s(e^{jw}) H(e^{jw}) \quad (2)$$

Speech signal can be further decomposed as convolution of residual over all pole filter. i.e.

$$s(t) = e(t)*a(t) \quad (3)$$

Grenze ID: 02.ICSIPCA.2017.1.36

$s(e^{jw}) = e(e^{jw})\, a(e^{jw})$ (4)

Substituting eq (4) in (2) gives,

$u(e^{jw}) = e(e^{jw})\, a(e^{jw})\, H(e^{jw})$ (5)

where $e(e^{jw})$ is the excitation signal, $a(e^{jw})$ is the all pole filter model of clean signal and $H(e^{jw})$ is the room transfer function.

II. Multi Channel Reverberant Speech Enhancement using LP Residual Cepstrum [1]

In the single channel speech dereverberation method proposed in [4], it is showed that lp co-efficient of reverberated speech is same as of clean speech. Hence eq (5) can be re written as below

$u(e^{jw}) = R(e^{jw})\, a(e^{jw})$ (6)

Where

$R(e^{jw}) = e(e^{jw})\, H(e^{jw})$ (7)

$R(e^{jw})$ is the Fourier transformation of reverberant prediction residual and $a(e^{jw})$ is the lp co-efficient of clean speech signal.

This approximation is observed due to the robustness of the LP coefficients under reverberation [4]. The robustness of LP coefficients to reverberation is illustrated in Figure 1 using LP spectrogram of clean speech and reverberated speech at direct to reverberation ratio (DRR) of -3dB. The spectrograms are computed from one sentence of the TIMIT database.

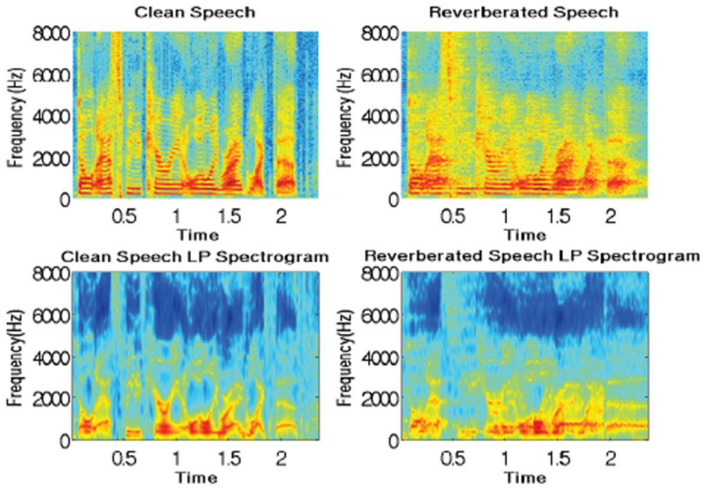


Figure 1. Comparison of the spectrograms of clean and reverberated speech. FFT spectrogram (Top row) and LP spectrogram (Bottom row)

In this context, it can be inferred that, if the clean speech residual is recovered from the reverberated speech residual, the dereverberated speech signal can be synthesized. The separation of clean residual from reverberated residual is performed using de-convolution. The de-convolution is performed using cepstral subtraction [5]. The cepstrum [6] of the reverberated residual is obtained and the peaks in higher quefrency of the cepstrum correspond to the AIR [7]. Hence peak picking is applied to the cepstrum of reverberated signal. The peaks obtained correspond to the cepstrum of AIR. The peaks are then subtracted from the reverberated residual signal so as to perform de-convolution and obtain an estimate of clean speech residual signal. The dereverberated signal is finally obtained by synthesizing of estimated clean speech residual signal and the LP coefficients of reverberated signal. Though the above said method will reduce the reverberation but not completely eliminate. This can be further improved using multichannel as show in fig 2.

In this method the single channel method of speech enhancement is used to perform de-convolution of AIR from reverberated residual signal at each microphone. The dereverberated output from each single microphone output are spatially filtered using a delay and sum beamformer (DSB) [8]. In order to eliminate the remaining spurious peaks a temporal averaging [9] method is used. The temporal averaging is applied on the LP residual of DSB output as shown in the Figure 2. The temporal averaging requires an accurate detection of glottal closure instants (GCI) which is computed using the dynamic programming projected

phase-slope algorithm (DYPSA) [10], [9], [11]. The DYPSA is preferred here because it is robust to reverberation. The advantage of the proposed method over spatio-temporal averaging method [9] is that, the DYPSA is applied on enhanced speech as explained in [4]. Hence, DYPSA is more accurate in the detection of GCI.

The proposed method performs reasonably better in terms of enhancing the reverberated signal as shown in below figure.

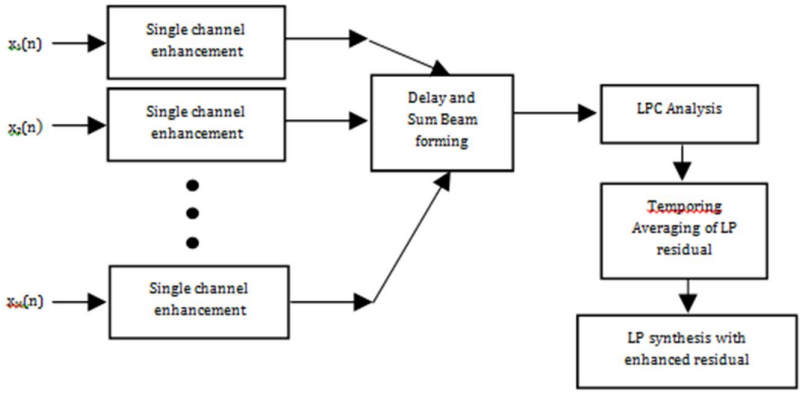


Figure. 2. Block diagram of the multi channel speech enhancement using LP residual cepstrum

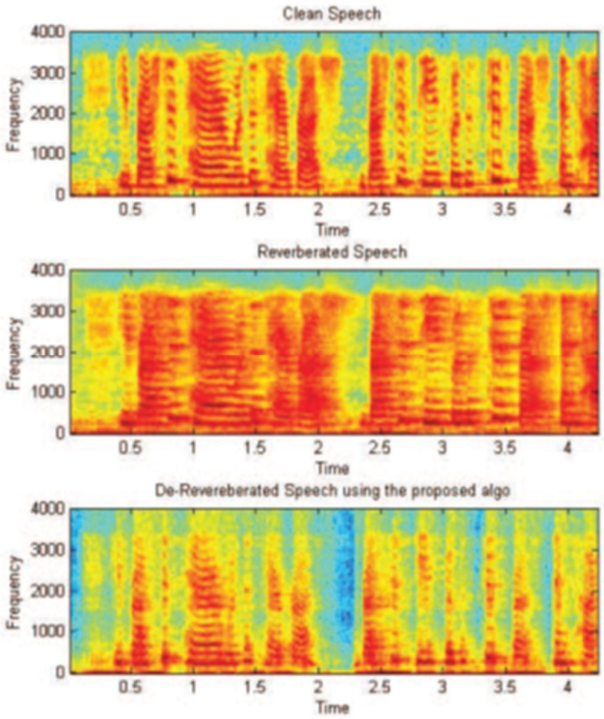


Figure3. Spectrograms for (a) Clean Speech (b) Reverberated Speech and (c) Dereverberated Speech

*Conclusion:*A multi channel speech enhancement method based on the LP residual cepstrum is proposed in this work. The de-convolution of acoustic impulse response from reverberated signal in each individual channel removes early reverberation. This dereverberated output from each channel is then spatially filtered using delay and sum beamformer. The late reverberation components are then removed by temporal averaging of the glottal closure instants (GCI) computed using the (DYPSA). The multi channel technique

performs better, when compared to spatio-temporal averaging alone. The method described in this work is computationally efficient compared to other conventional de-convolution methods which often rely on the estimation of the acoustic impulse response. However, the peak detection algorithm is prone to errors at very low DRR. This situation is often encountered in large rooms due to spurious peaks present in the regions, where late reverberation components exist. The performance of the proposed method needs to be further investigated for long reverberation times and low signal to noise ratios.

III. MULTI-CHANNEL SPEECH DEREVRBERATION BASED ON A STATISTICAL MODEL OF LATE REVERBERATION [2]

Short time power spectral density (STPSD) of reverberant signal is,

$\gamma_{xx}(t, f) = \gamma_{xdxd}(t, f) + \gamma_{xrxr}(t, f),$ (8)

$\gamma_{xrxr}(t, f) = e^{-2\alpha T} \gamma_{xx}(t, f)\ (t - T, f).$ (9)

Where,

$\boldsymbol{\gamma_{xx}}$ is the STPSD of reverberant signal,

$\boldsymbol{\gamma_{xdxd}}$ is the STPSD of direct signal

$\boldsymbol{\gamma_{xrxr}}$ is the STPSD of later echo signal

$\mathbf{T_s < T << T_r}$ is reverberant time of the room

$\mathbf{T_s}$ is the time span over which the speech signal can be considered stationary, which is usually around 20-40 m

$\mathbf{T_r}$ is reverberation time of room

$\boldsymbol{\alpha}$ is a constant. linked to reverberation time T_r

Numerous techniques for the enhancement of noisy speech degraded with uncorrelated additive noise have been proposed in literature. Among them the spectral subtraction methods are the most widely used due to the simplicity of implementation and the low computational load, which makes them the primary choice for real-time applications. A common feature of this technique is that the noise reduction process can be related to the estimation of a Short-Time Spectral Attenuation factor. Since the spectral components are assumed to be statistical independent, this factor is adjusted individually as a function of the relative local A Posteriori Signal to Noise Ratio on each frequency. The A Posteriori SNR is defined as

$$SNR_{post}(t,f) \cong \frac{|X(t,f)|^2}{\gamma_{x_r x_r}(t,f)}$$

The estimate of the amplitude spectrum of the noise is given by

$$\left|\tilde{S}(t,f)\right| = G(t,f)\,|X(t,f)|$$

Where G(t,f) is the gain function given as,

$$G(t,f) = 1 - \frac{1}{\sqrt{SNR_{post}(t,f)}}$$

In all frames it is however possible that for some frequencies the estimated amplitude of the noise spectrum is larger than the instantaneous amplitude of the noisy speech spectrum |X(t,f)|. Since this could lead to negative estimates for the amplitude of the clean speech spectrum $|\hat{}\ S(t, f)|$, for these frequencies the gain function G(t, f) is usually put to zero (i.e. half-wave rectification) or equal to a small noise floor value say λ as proposed in [12] results in the following gain function,

$$G(t,f) = \begin{cases} 1 - \frac{1}{\sqrt{SNR_{post}(t,f)}} & \text{if } \left|\tilde{S}(t,f)\right| \geq \lambda |X(t,f)| \\ \lambda & \text{Otherwise} \end{cases}$$

For single-channel noise reduction additional effort has to be made to reduce residual noise which is mainly caused by the random variations due to the reverberation in |X(t, f)|. Under the assumption that the speech

signals are time aligned it can be shown that in the multi-channel case this variance can be reduced by replacing the amplitude spectrum |X(t, f)| by a spatially averaged value, i.e.

$$\left|\overline{X(t,f)}\right| = \frac{1}{N}\sum_{n=0}^{N-1}\left|X_n(t,f)\right|$$

where N denotes the number of microphones.
Experimental setup and implementation overview of algorithm is as shown in fig 4 and 5 below.

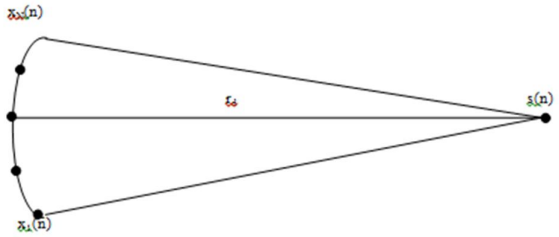


Figure 4. Experimental setup

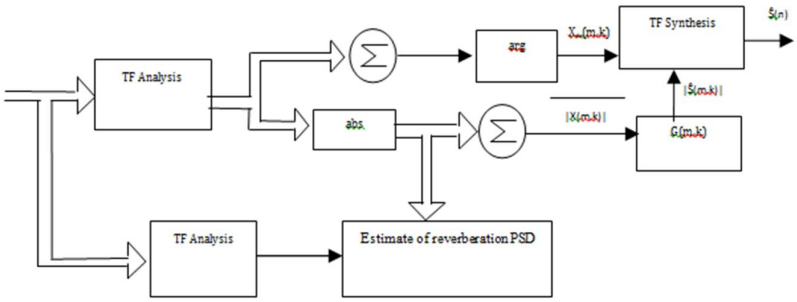


Figure 5. Overview of algorithm

Conclusion: In this method it is shown how multiple microphone signals can be used to obtain an accurate estimate of the power spectrum of the late reverberant signal. Experimental results show a decrease in reverberation and distortion when using more microphones. Additionally, the fine structure of the speech signal is partially restored due to spatial averaging. Future work will focus on more accurate modelling of the RIR, loosening the assumptions w.r.t. the geometry of the microphone array and application in a real acoustic environment, rather than a simulated one.

IV. BLIND DEREVERBERATION ALGORITHM FOR SPEECH SIGNALS BASED ON MULTI-CHANNEL LINEAR PREDICTION [3]

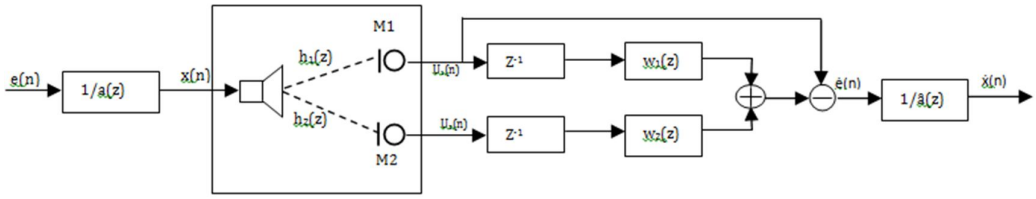
As shown in eq (5) that reverberated signal can be expressed as residual signal $e(e^{jw})$ convolved with $a(e^{jw})$ and $H(e^{jw})$. If we can extract $e(e^{jw})$ and $a(e^{jw})$ using prediction filter $a(e^{jw})$, then clean speech can be extracted as shown in fig 6
Although the developments are presented for the particular case of two microphones, the method could be extended to multi-microphone.
The algorithm is constructed with the following hypotheses:

- It is assumed that input signal x(n) is generated from a finite AR process applied on white noise e(n) The AR polynomial is
  a(z) = 1 - {a1z-1 + ... aNz-N } (10)

- It is assumed that room transfer functions $H_1(z)$ and $H_2(z)$, modelled by polynomials, are time-invariant and have no common zeros.

$$Hi(z) = \sum_{k=0}^{m} hi(k)Z^{-k} \qquad (11)$$

Let us call the signals received at the microphones M1and M2, $u_1(n)$ and $u_2(n)$ respectively. They are obtained by filtering x(n) with the room transfer function



Figue 6 Schematic diagram of dereverberation system to recover the input signal from the microphone signals

Prediction Filter:

The main task of this algorithm is estimation of prediction filter and lp co-efficient. Looking at fig 6 prediction error ê(n) can be expressed as

$$\hat{e}(n) = u_1(n) - (w_1(n)*u_1(n-1)+w_2(n)*u_2(n-1)) \qquad (12)$$

Eq (12) can be written as,

$$\hat{e}(n) = x_n^T h_1 - x_{n-1}^T Hw \qquad (13)$$

Where,

$x_n = \{ x(n), x(n-1) ,..., x(n-(m+L)) \}^T$

$h_1 = \{h_{1,0} ,..., h_{1,m},0,0,...,0\}^T$

H is a full row-rank matrix of size (m + L) * 2L and

$2L \geq m + L$

H = [H1,H2],

Hi is a (m + L) * L convolution matrix expressed as

$$\boldsymbol{H}_i = \begin{pmatrix} h_{i,0} & 0 & \dots & 0 \\ h_{i,1} & h_{i,0} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \\ h_{i,m} & & & 0 \\ 0 & h_{i,m} & & h_{i,o} \\ \vdots & & \ddots & \vdots \\ 0 & \dots & 0 & h_{i,m} \end{pmatrix}, \quad i = 1, 2,$$

w is the prediction filter set,

$w = [w^T_1 w^T_2]^T$, &

$w_i = [w_{i;0}; \dots ; w_{i;L-1}]^T$,

Minimizing the mean square value of the prediction error gives us:

$$w = (H^T E\{x_{n-1}x^T_{n-1}\} H)^+ H^T E\{x_{n-1}x^T_n \} h_1 \qquad (14)$$

where A^+ is the Moore-Penrose generalized inverse of matrix A [15], and E{ } is an expectation operator. If we replace the column vector h1 with matrix H, we can define matrix Q as:

$$Q = (H^T E\{x_{n-1}x^T_{n-1}\} H)^+ H^T E\{x_{n-1}x^T_n \} H \qquad (15)$$

As the input signal can be generated by an AR process we can write

$x_n = C^T x_{n-1} + e_n$

Where

C is the companion matrix of lp co-efficient

$e_n = [e(n),0,....,0]^T$

Then we can write

$E\{x_{n-1}x^T_n\} = E\{x_{n-1}x^T_{n-1}\} C$ (16)

Assuming that $E\{x_{n-1}x^T_{n-1}\}$ is positive definite, we can replace it with X^TX where X is a matrix. Matrix Q is thus expressed as:

$Q = H^T(HH^T)^{-1}CH$ (17)

or

$w = H^T(HH^T)^{-1}Ch_1$ (18)

Replace **w** in eq 13

$\hat{e}(n) = x_n^T h_1 - x_{n-1}^T HH^T(HH^T)^{-1}Ch_1$

$\hat{e}(n) = x_n^T h_1 - x_{n-1}^T Ch_1$

$= (x_n^T - x_{n-1}^T C)h_1$

$= e_n^T h_1$

$= h_{1,0}e(n)$ (19)

The above equation shows prediction error is proportional to white noise e(n) or clean speech residual error.

Calculation of matrix Q:

Matrix Q can be calculated with the signals received at the microphones. Using the matrix notation defined previously, the microphone signals can be expressed as:

$u_n = H^T x_n$ (20)

where $u_n = [u_1(n) \ldots u_1(n-L); u_2(n) \ldots u_2(n-L)]^T$. Using relation (15) and (20), we can express matrix Q as a function of the microphone signals:

$Q = E(un\text{-}1uTn\text{-}1\} + E\{un\text{-}1uTn\}$ (21)

Equation (21) is used in practice to calculate Q.

Estimated AR process:

Determinant of companion matrix C is

$\lambda(C) = -\lambda^N\{1-(a_1\lambda^{-1}+\ldots+ a_N\lambda^{-N}\}$ (22)

Let us consider the non-zero Eigen values of matrix Q [13]

$\lambda(Q) = \lambda(H^T(HH^T)^{-1}CH)$

$\lambda(Q) = \lambda(HH^T(HH^T)^{-1}C)$

$\lambda(Q) = \lambda(C)$ (23)

From Eq. (16) we deduce that the estimated AR polynomial, $\hat{a}(z)$, can be obtained from the characteristic polynomial of matrix Q.

*Conclusion:*The method enables the precise recovery of speech signals suffering from room reverberation. In particular, the output signal is not whitened as found with traditional dereverberation techniques. The excellent simulation results show the potential of the method and prove its solid theoretical background. However, the current method suffers from several limitations. First, we are currently limited to short room impulse responses. Indeed, a longer room impulse response would require longer prediction filters and thus a larger matrix Q. In this case, computational time and accuracy would become an issue. One major reason for this problem may be that the two transfer functions have numerically common zeros. Moreover, the current results were obtained for a noise free environment, which is quite unrealistic. However, in theory if the hypotheses are satisfied, the method could be extended. Future work will thus consist in improving the method to cope with longer room impulse responses and noisy environments.

V. CONCLUSION

All three paper shows that increase in microphone will give better result against single microphone for the given methods at the cost of increase in computational complexity. Computationally blind dereverberation [1] using LP residual cepstrum is simple. Its performance will be better if DRR is large and degrades as the DRR decreases (or room size increases).

Speech dereverberation using statistical model [2] gives better performance in low DRR and also computational complexity is simple. Its drawback is it works on predetermined reverberation time T_r. That means for a given room we need to estimate T_r and feed it to algorithm.

Speech dereverberation based on linear prediction[3] gives very good result compared to other two but its computational complexity is very high. Finding the Q for large room in real time environment is difficult to achieve.

REFERENCES

[1] Harish Padaki, Karan Nathwani, and Rajesh M Hegde, "Multi channel speech dereverberation using the lp residual cepstrum," in *Communications(NCC), 2013 IEEE National Conference*. pp. 1–5.

[2] E.A.P. Habetts, " Multi-channel speech derevrberation based on a statistical model of late reverberation in 2005 Accostic, speech and signal processing (ICASSP) IEEE International conference,

[3] M. Delcroix, T.Hikichi, & M Miyoshi, "Blind de reverberation algorithm for speech signals based on multi-channel linear prediction," Acoust. Sci. Technol., vol. 26, no. 5, pp. 432–439, 2005.

[4] Harish Padaki, Karan Nathwani, and Rajesh M Hegde, "Single channel speech dereverberation using the lp residual cepstrum," in Communications(NCC), 2013 National Conference on. IEEE, 2013, pp. 1–5.

[5] K. Furuya, S. Sakauchi, and A. Kataoka, "Speech dereverberation by combining mint-based blind deconvolution and modified spectral subtraction," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*. IEEE, 2006, vol. 1, pp. I–I.

[6] S. Xizhong and M. Guang, "Complex cepstrum based single channel speech dereverberation," in *Computer Science & Education, 2009. ICCSE'09. 4th International Conference on*. IEEE, 2009, pp. 7–11.

[7] D. Bees, M. Blostein, and P. Kabal, "Reverberant speech enhancement using cepstral processing," in *Acoustics, Speech, and Signal Processing,1991. ICASSP-91., 1991 International Conference on*. IEEE, 1991, pp. 977–980.

[8] Jacob Benesty, Jingdong Chen, and Yiteng Huang, *Microphone arraysignal processing*, vol. 1, Springer, 2008.

[9] Patrick A Naylor, Anastasis Kounoudes, Jon Gudnason, and Mike Brookes, "Estimation of glottal closure instants in voiced speech using the dypsa algorithm," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 1, pp. 34–43, 2007.

[10] Anastasis Kounoudes, Patrick A Naylor, and Mike Brookes, "The dypsa algorithm for estimation of glottal closure instants in voiced speech," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*. IEEE, 2002, vol. 1, pp. I–349.

[11] Mark RP Thomas, Jon Gudnason, and Patrick A Naylor, "Estimation of glottal closing and opening instants in voiced speech using the yaga algorithm," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 82–91, 2012.

[12] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," IEEE ICASSP'79, vol. 4, pp. 208–211, 1979.

[13] D. A. Harville, Matrix Algebra from a Statistician's Perspective (Springer-Verlag, New York, 1997).