

Cancellation of noise from speech using Kepstrum analysis

T.J.Moir

IIMS, Massey University at Albany, Auckland, New Zealand

t.j.moir@massey.ac.nz

Abstract

A method for canceling stationary or non-stationary noise from speech is presented. The technique uses two microphones a short distance apart and an estimate of the acoustic transfer function between the microphones when speech is absent. A voice activity detector (VAD) is used based on a spatial 'active zone' within which all signals are assumed to be desired signals and out-with which all signals are considered noise. A measure of time-difference of arrival (TDOA) is used together with a transfer function estimate based on Kepstrum analysis.

1.Introduction

The problem of suppression of noise from speech has applications in such areas as speech recognition and hands-free telephony in automobiles. The basic idea has been around for some time [1] and normally involves the use of at least two microphones. One microphone is placed near the noise source and the second microphone is placed such as to pick up the speech and the noise. An adaptive filter based on the least-mean-squares (LMS) method is then used to minimise the mean-squared-error and arrive at an adaptive tracking Wiener filter. It is well known however that in order for such a system to work effectively requires good coherence between the two microphones and this necessitates the microphones not being too far apart. Conversely if the microphones are too close together then the speech as well as the noise will be cancelled. Although there are certain applications in special environments which are suitable to such a solution it is generally accepted that modifications of this fundamental idea is necessary in many real environments. There are quite successful approaches which use microphone arrays for instance [2,3] for noise cancellation and for spatial filtering or beam-forming[4]. The approach used here wishes to simplify the procedure as much as possible for future real-time implementation and so two microphones are used. The basic idea used herein has already been implemented using the LMS algorithm[5]. However it was found that although good improvements in signal to noise ratio (SNR) could be claimed in theory and simulation, that for real environments any improvements were marginal being at most 3dB. In order to improve the SNR a multiple sub-band approach was then used and this did give rise to realistic improvements in SNR but at the expense of a significant increase in computation.[6]

The method used here relies almost entirely on the fast-Fourier transform (FFT) for both the estimation of time-delay of arrival (TDOA) and for the identification of the acoustic-path transfer functions. When the microphone sensors are close together rather than far apart, a voice activity detector (VAD) is required in order to determine times when speech is absent (when identification of the acoustic path transfer functions are carried out). Fortunately a robust method for VAD has already been given by Agaiby and Moir [7] which also uses time-delay estimation. Clearly the same algorithm used for TDOA can then be used in the VAD making the entire algorithm based on FFTs with the exception of a single finite-impulse response (FIR) convolution. The convolution can of course be implemented in the frequency domain using FFTs but this is only more computationally efficient when the FIR filter order becomes greater than about 50.

2.Noise cancellation basic idea

Single noise source – no signal.

Suppose we have a single noise source and two microphones receiving this noise source. Assume the microphones are relatively close together (say 10-20cm apart). Then the noise source will reach each microphone via different acoustic path FIR transfer functions as in the figure below.

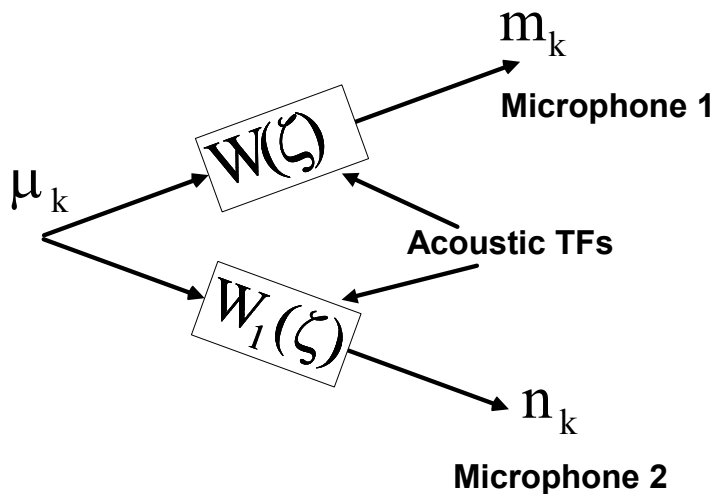


Figure 1 Single noise source and acoustic path transfer functions

To simplify the mathematical notation it is assumed that the two transfer functions (TFs) are function of $\zeta = z^{-1}$ where z is the complex z -transform operator. It is important to note that that $W(\zeta) \neq W_1(\zeta)$ i.e. that the noise source cannot be directly behind or in front of the two microphones. To further simplify the notation the argument ζ is often dropped altogether so that the signals received at the two microphones can be written as

$$m_k = W\mu_k \quad (1)$$

and

$$n_k = W_1\mu_k \quad (2)$$

where μ_k is a noise source. From (1) and (2) above we can write

$$m_k = \frac{W}{W_1}n_k \quad (3)$$

or

$$m_k = Hn_k \quad (4)$$

where $H = W/W_1$ is the transfer function *between* the two microphones. Let us now suppose that somehow we can estimate this transfer function H . It is assumed that H is causal and if this is not the case then the two microphones can be reversed. Practically this means that any pure time-delay present in W_1

must be smaller than that in W . Furthermore H is an FIR transfer function and so it must also be stable. Normally H will include a net time-delay say d which will be the difference of the individual delays in each FIR transfer function normally termed the time-delay of arrival (TDOA). Now if H is included as a filter after n_k , then we can create an error signal.

$$e_k = m_k - Hn_k \quad (5)$$

Single noise source with signal.

The signal here is assumed to be of an intermittent nature (speech) and it is assumed that during silence periods or pauses in the speech that the above transfer function H can be estimated. During periods of speech, s_k is assumed to be added directly to the two microphones with negligible transfer function (as the speaker is close to the two microphones).

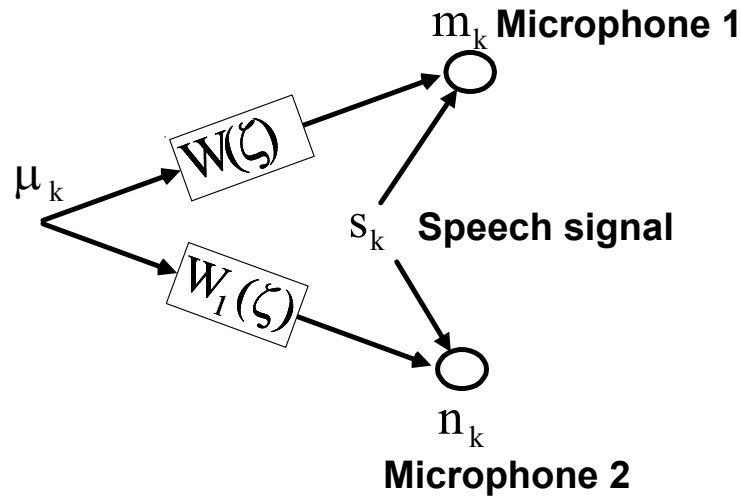


Figure 2 Single noise source with speech signal

We must now have.

$$m_k = W\mu_k + s_k \quad (6)$$

and

$$n_k = W_1\mu_k + s_k \quad (7)$$

The error now becomes

$$e_k = m_k - Hn_k \quad (8)$$

$$= W\mu_k + s_k - H(W_1\mu_k + s_k) \quad (9)$$

and using the fact that $H = \frac{W}{W_1}$ the error signal becomes

$$e_k = (1 - \frac{W}{W_1})s_k \tag{10}$$

which is a filtered version of the speech signal and the noise has been cancelled. The problem is now one of being able to estimate the FIR transfer function H during periods of no speech. The complete diagram is shown below.

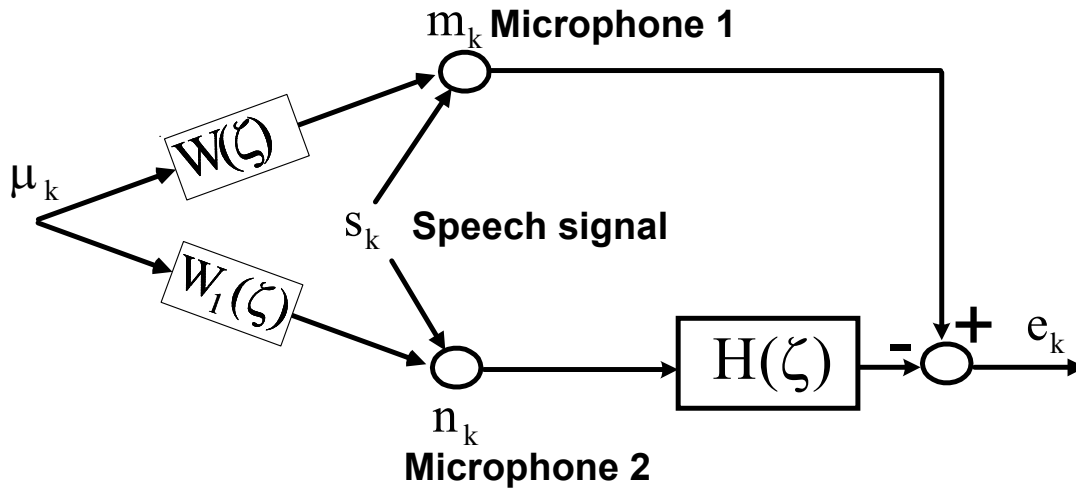


Figure 3 Single noise source with speech signal and error signal.

As previously it is assumed that the noise source does not appear directly in front of or behind the two microphones.

3.The voice activity detector (VAD)

In the previous section, the only way to identify the transfer function H is when the speech signal is absent and this requires a VAD. A convenient and robust method for VAD is given in [7]. It is assumed that the desired speech is spoken directly in front of the two microphones so that the time-delay of arrival (TDOA) will be close to zero. Similarly any noise source will have a TDOA which is much greater than zero. By restricting the area in which valid speech is assumed to be formed (a zone of speech activity), simply by estimating the TDOA and ensuring that this delay is greater than a pre-defined threshold (say d_{max}) will determine whether any signal arriving at the microphones is due to signal or noise. Clearly the VAD is working on the geometry of the problem rather than the more usual approaches using energy thresholds. It was shown [8] that the zone of speech activity is a two sheet hyperboloid which extends in front as well as behind the two microphones as shown in Fig 4 below.

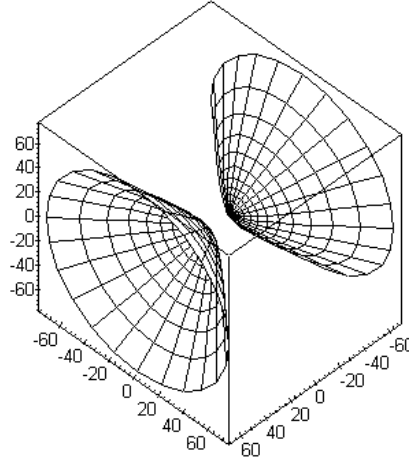


Figure 4 Zone of active speech when microphones are 20cm apart (two sheet hyperboloid. Units for all axis in cm)

The TDOA between the signals received at both microphones m and n could be estimated using the standard cross-correlation method. However it is known that for robust estimation of time-delay that the signals received at the two microphones must be white. When the signals are white a prominent peak is produced in the cross-correlation. For our problem the signals are rarely if ever white and so instead the generalized cross-correlation method (GCC) [9] method is used instead. The GCC method corresponds to a maximum likelihood estimate of the time delay. The GCC algorithm is usually described in continuous time as the time τ at which the GCC function $R_{nm}^g(\tau)$ is maximum where

$$R_{nm}^g(\tau) = \int_{-\infty}^{\infty} \Psi_g(f) G_{nm}(f) e^{j2\pi f\tau} df \quad (11)$$

$G_{nm}(f)$ is cross-spectral density between the two microphones n and m and we define

$$\Psi_g(f) = \frac{|\gamma_{nm}(f)|^2}{|G_{nn}(f)|[1-|\gamma_{nm}(f)|^2]} \quad (12)$$

$|\gamma_{nm}(f)|^2$ is magnitude-squared coherence (MSC) between n and m defined as

$$|\gamma_{nm}(f)|^2 = \frac{|G_{nm}(f)|^2}{G_{nn}(f)G_{mm}(f)} \quad (13)$$

In (13) above $G_{nn}(f)$ and $G_{mm}(f)$ are the respective spectral densities of the received signals at the microphones n and m .

Valid speech is assumed when

$$\text{Estimated time delay} \leq T_{\max} \quad (14)$$

and

$$\text{average MSC} \geq C_{\min} \quad (15)$$

T_{\max} is chosen depending on how narrow the zone of active speech is to be (normally about 4-5 samples in discrete time sampled at 22050kHz) whilst C_{\min} is chosen to be around 0.5. Equation (15) is necessary because of reverberations. It is known for instance that reverberant speech has a lower MSC than non-reverberant speech and so this goes towards reducing false readings. Practically C_{\min} acts a bit like a sensitivity control. Too high and the time-delay speed of response is sluggish. Too low and the response is too sensitive to reverberations. The average MSC is taken over all frequencies (up to half sampling in discrete-time). The discrete-time version of this algorithm is described later in the paper.

4. Identification using the Kepstrum method.

A convenient method of identifying the acoustic transfer functions $W(\zeta)$ and $W_1(\zeta)$ is by using Complex Cepstrum or Kepstrum [10]. The procedure is computationally quite simple and uses mostly the fast-Fourier transform (FFT) and its inverse (IFFT).

Without loss of generality we assume that $W(\zeta)$ and $W_1(\zeta)$ are finite-impulse response (FIR) transfer functions.

The procedure is carried out during periods of non-speech (or noise-alone). First the periodogram is estimated from the FFT. The periodogram is a discrete estimate of the continuous-time spectral-density. Consider N samples per frame and a time-domain noise data vector given for one of the microphones as

$$x_k = [n_0, n_1, \dots, n_{N-1}]^T \quad (16)$$

The periodogram is the estimate of spectral density at each frequency-bin j and is found from

$$\hat{S}_j = \frac{1}{N} |X_j|^2, \quad j = 0, 1, 2, \dots, N-1 \quad (17)$$

Where $X_j = F(x_k)$ and F denotes FFT. After taking the natural log of periodogram it is found that there exists a bias equal in magnitude to minus Euler's constant $\gamma = 0.577215\dots$ [11]. Here the symbol for Euler's constant should not be confused with coherence which also uses a similar symbol in the literature. The Kepstrum coefficients are found from the inverse of the natural logarithm of the periodogram

$$k_i = F^{-1}(\log(\hat{S}_i) + \gamma), \quad i = 0, 1, 2, \dots, N-1 \quad (18)$$

Where the N Kepstrum coefficients satisfy the symmetry $k_i = k_{N-i}, i = 1, 2, \dots, \frac{N}{2} - 1$. The zeroth

Kepstrum coefficient must be halved thus $k_0 = k_0 / 2$. Hence for two sets of Kepstrum coefficients $k_i, k'_i, i = 0, 1, \dots, N$, one set from each microphone it is easily shown by Sylvia and Robinson [12] that

the corresponding impulse response of length $\ell \leq \frac{N}{2} - 1$ is found from

$$h_0 = \exp(k_0 - k'_0) \quad (19)$$

and the recursion

$$(n + 1)h_{n+1} = \sum_{m=0}^n (n + 1 - m)h_m (k_{n+1-m} - k'_{n+1-m}), n = 0, 1, 2, \dots, \ell - 1 \quad (20)$$

where

$$H(\zeta) = h_0 + h_1\zeta + h_2\zeta^2 + \dots + h_\ell\zeta^\ell \quad (21)$$

The recursion (20) being in fact a convolution. The difficulty with this approach is however that the Kepstrum method will not detect time-delays as no cross-spectral phase information is available. Whilst it may be possible to modify the Kepstrum method to include phase, it is simpler to use the estimate of phase already made in the VAD using the GCC method. For example if the time-delay is estimated to be d steps, then the FIR transfer function H in (21) need only be shifted by d coefficients to the right to account for the delay.

5. Discrete algorithm for Kepstrum noise-cancellation.

The previous theory gives rise to a noise-canceller which uses the GCC method for estimation of time-delay. This time-delay estimate is used in both the discrete VAD and to modify the FIR transfer function H as the Kepstrum method cannot estimate time-delay of arrival. A block diagram of the overall method is shown in Figure 5 below. The error signal is not shown but found from (5) using convolution.

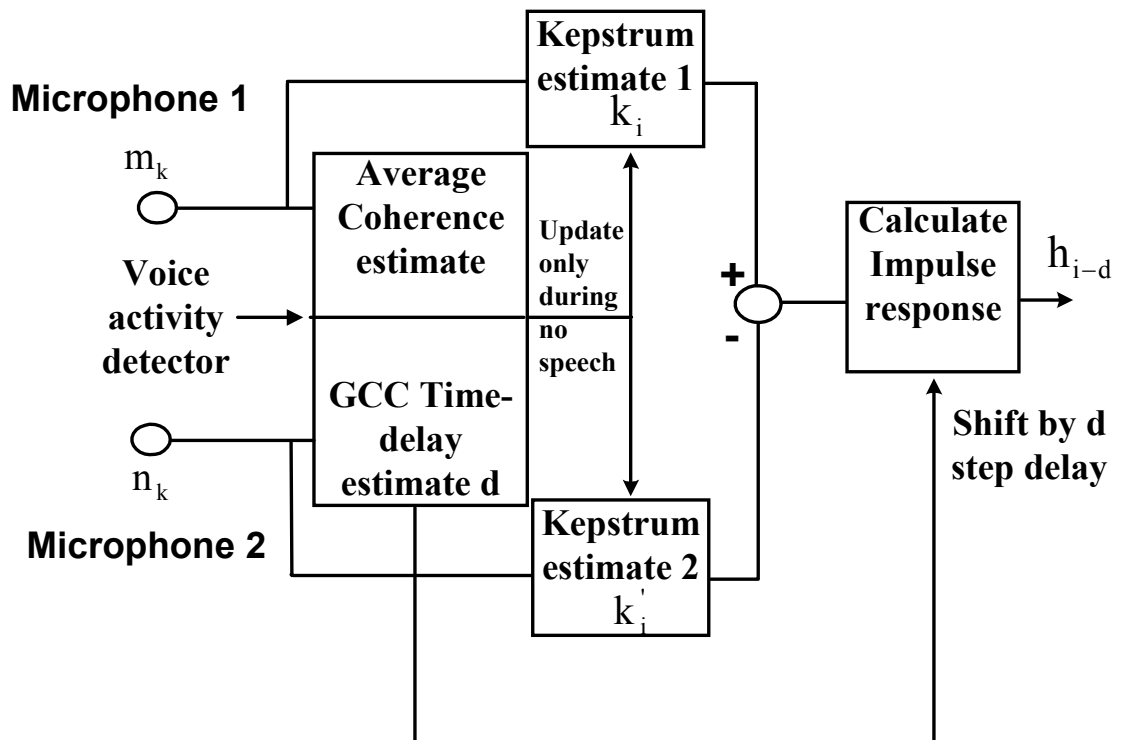


Figure 5 Block diagram of Kepstrum noise-canceller.

Algorithm 5.1: Discrete-time VAD

Step 1: At each FFT frame index $i=1,2,3,\dots$ assign the two vectors

$$x_k = [n_0, n_1, \dots, n_{N-1}]^T$$

and

$$y_k = [m_0, m_1, \dots, m_{N-1}]^T$$

which are composed of N samples of the two microphone inputs and have been suitably windowed with their corresponding frequency vectors corresponding to X_j and Y_j respectively.

Estimate the spectra (periodograms) of the signals from each of the two microphones

$$\hat{S}_{nn}(i) = \beta \hat{S}(i-1) + (1-\beta)XX^* \quad (22)$$

$$\hat{S}_{mm}(i) = \beta \hat{S}(i-1) + (1-\beta)YY^* \quad (23)$$

(22) and (23) is a method of smoothly updating the spectrum recursively at each FFT frame rather than the straight batch method of (17). In the above equation $*$ represents complex conjugate and $0 \leq \beta < 1$ is a forgetting factor. For the results used in this paper $\beta = 0.5$ was used as a compromise between fast tracking and smoothing. If chosen to be too large then the tracking ability of the GCC time-delay estimator is severely compromised. Some experimentation is required depending on the application.

Estimate the cross-spectrum (cross-periodogram) from

$$\hat{S}_{nm}(i) = \beta \hat{S}(i-1) + (1-\beta)XY^* \quad (24)$$

Step2 :Estimate the MSC at each FFT frame from

$$|\hat{\gamma}_{nm}(i)|^2 = \frac{|\hat{S}_{nm}(i)|^2}{\hat{S}_{nn}(i)\hat{S}_{mm}(i)} \quad (25)$$

and at each frame i , average over frequency k the MSC thus

$$|\bar{\gamma}_{nm}(i)|^2 = \sum_k |\hat{\gamma}_{nm}(i)|^2 \quad (26)$$

Step 3: Estimate the term $\psi_g(i)$ from

$$\psi_g(i) = \frac{|\hat{\gamma}_{nm}(i)|^2}{|\hat{S}_{nn}(i)|[1-|\hat{\gamma}_{nm}(i)|^2]} \quad (27)$$

Step 4: Estimate the time-delay of arrival d from the generalised cross-correlation

$$\hat{R}_{nm}^g(d) = \max F^{-1} \{ \Psi(i)\hat{S}_{nm}(i) \} \quad (28)$$

That is, the maximum of the inverse FFT of $\Psi(i)\hat{S}_{nm}(i)$. A positive delay can be inferred if the maximum occurs in the region $0 < d < N/2 - 1$ ie the first half of the inverse FFT and a negative delay if the maximum occurs in the upper half of the inverse FFT. This is important for the noise cancellation algorithm as the ratio of the two acoustic transfer functions H must always be causal.

Valid speech is then assumed when

$$\text{Estimated time delay} \leq d_{\max} \quad (29)$$

and

$$|\bar{\gamma}_{nm}(i)|^2 \geq C_{\min} \quad (30)$$

For the experiments carried out in this paper a sampling interval of 22050Hz was used so that each sample interval corresponds to $45.35 \mu\text{s}$. Typically d_{\max} was chosen to be no more than 5 samples and C_{\min} was chosen as 0.5. This algorithm is used as the front-end to the main noise-cancellation algorithm which follows.

Algorithm 5.2: Kepstrum noise-cancellation algorithm

Initialise the coefficients of the FIR filter H to some nominal values.

Step 1: At each FFT frame index $i=1,2,3,\dots$ determine using Algorithm 5.1 if the sampled waveforms from the two microphones correspond to valid speech plus noise or to a period of 'noise-alone'. Also using Algorithm 5.1 determine the TDOA d .

Step 2: If $d < 0$ then swap the signals from the microphones.

Step 3: If valid speech plus noise then freeze the coefficients of the transfer function H and go to step 4 else estimate the two sets of Kepstrum coefficients $k_i, k'_i, i = 0, 1 \dots N$ from the 'noise-alone' period.

Step 4: Compute the coefficients of the FIR filter H using (19) and (20) and shift the coefficients d steps to the right.

Step 5: Compute the noise-cancelled output (error signal) $e(k)$ from (5).

6. Illustrative example

Consider a speech signal s_k corrupted by noise μ_k with the two composite signals received at each microphone being $m_k = s_k + 0.9\mu_{k-20}$ and $n_k = s_k + \mu_{k-10}$. The noise signal chosen was a competing talker. The SNR was measured for each channel was -4.8dB and -5.7dB. The TDOA is clearly 10 sample intervals or 0.453ms. The two channels were sampled at 22050Hz, 8 bits and an FFT frame length of $N=1024$ points was used. The forgetting factor for computing the periodogram and cross-

periodogram was chosen as $\beta = 0.5$ and the threshold for coherence was chosen as $C_{\min} = 0.5$. The zone of activity or active zone time-delay threshold was chosen as $d_{\max} = 5$ or 0.226ms. Figure 6 (top) shows the original desired speech signal. Figure 6 (middle) is the composite mix of the original speech signal and the competing talker whilst Fig 6 (bottom) is the enhanced speech signal together with the VAD flag. The segmented SNR of the enhanced speech signal was found to be approximately 17.9dB indicating a worst case improvement of 22.7dB. The enhanced speech had traces of the second speaker but these were momentary bursts at very low volume.

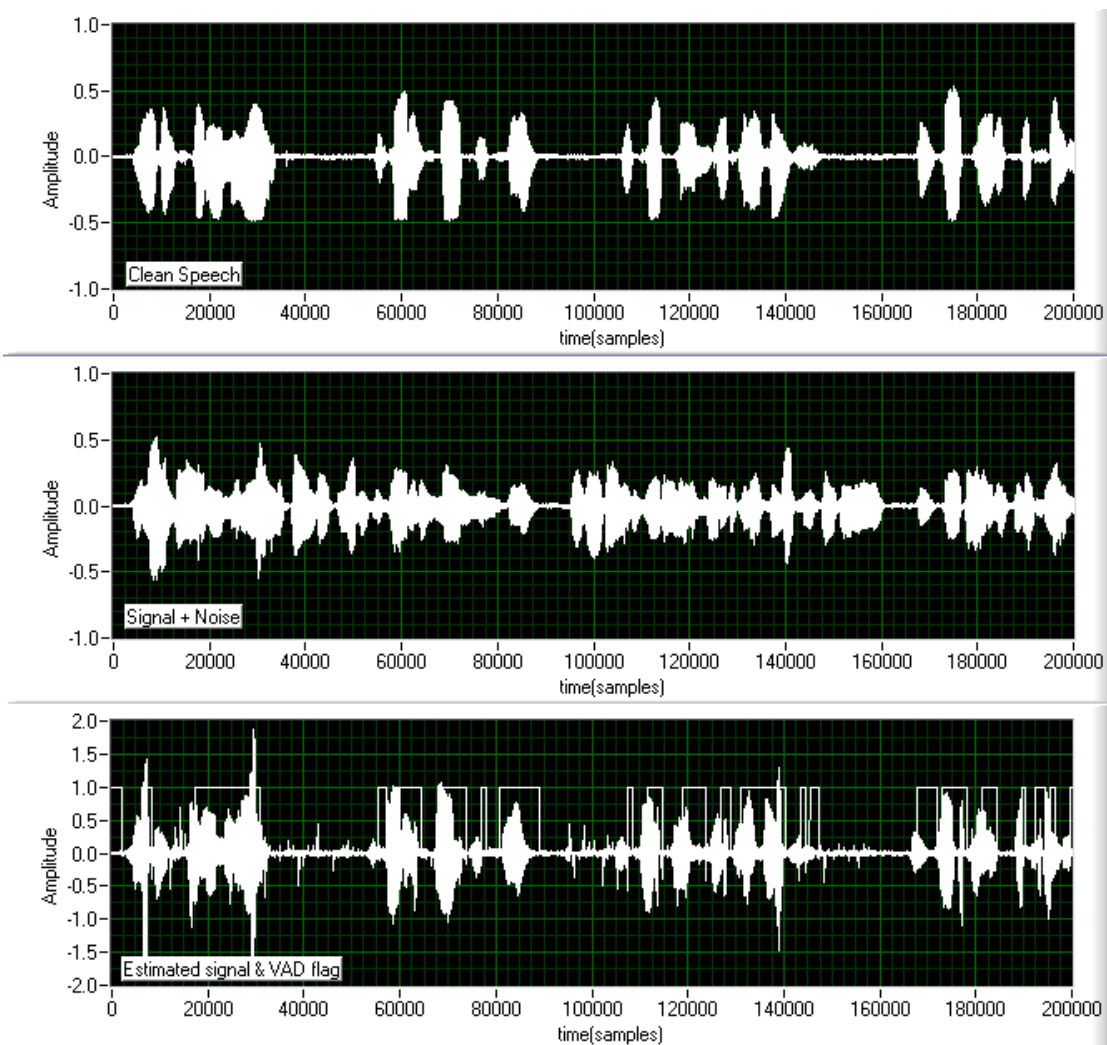


Figure 6. Clean speech (top) , corrupted speech (middle) and enhanced speech with VAD flag (bottom).

7. Conclusions

An approach to noise-cancellation with application to speech processing has been shown to give promising results. The method uses a time-delay voice-activity detector and Kepstrum system identification and is almost entirely reliant on the FFT. This makes the algorithm suitable for many real-time applications where the computational burden must be kept at a minimum. However, comparisons need to be made with other existing algorithms in real environments to realise the full potential of the work.

8. References

- [1] Widrow, B & Stearns, S.D. 1985 *Adaptive Signal Processing*. Prentice-Hall, Englewood Cliffs, NJ
- [2] Van Compernelle, D., Ma, W and Van Diest, M. 1989 Speech recognition in noisy environments with the aid of microphone arrays, *European Conference on Speech technology*, pp 757-660
- [3] Stadler, R.W and Rabinowitz, W.M 1993. On the potential of fixed arrays for hearing aids, *J Acoust. Soc. Am.* pp 1332-1342
- [4] McCowan, I A, Moore, D C and Sridharan, S. 2002 Near-field adaptive beamformer for robust speech recognition, *Digital Signal Processing* 12,1, pp87-106
- [5] Campbell, D R, Moir, T.J, and Dabis, H.S 1992 Multivariable polynomial matrix formulation of adaptive noise cancelling *Signal Processing*, 26, pp 177-183
- [6] Shields, P W and Campbell D R 2001 Improvements in intelligibility of noisy reverberant speech using a binaural subband adaptive noise-cancellation scheme *J Acoust. Soc Am.* 110, 6 pp 3232-3242
- [7] Agaiby, H & Moir, T.J. 1997 Knowing the wheat from the weeds in noisy speech, *Proc. 5th European Conf. Speech Comm. & Technology*, Rhodes, Greece, Sept 22-25.
- [8] Agaiby, H 1999 Word boundary detection for engineering applications, *PhD Thesis, University of Paisley, Scotland*
- [9] Knapp, C H and Carter, G C 1976 The generalized correlation method for estimation of time-delay, *IEEE Trans. Acoust., Speech, Signal Proc.*, ASSP-24,4, pp 320-327
- [10] Barrett, J.F. & Moir, T.J 1986. The Kepstrum method for spectral analysis. *Intern. J. Control*, vol.43, no.1, pp 29-57
- [11] Wahba, G. 1980 Automatic smoothing of the log periodogram, *J. Amer. stat. Assoc.*, vol. 75, pp 122-132.
- [12] Silvia, M.T & Robinson, E.A. 1978 Use of the Kepstrum in signal analysis. *Geoexploration*, vol.16, pp 55-73.