AN APPROACH TO DOUBLETALK DETECTION BASED ON NON-NEGATIVE MATRIX FACTORIZATION

Niall Cahill and Robert Lawlor

Department of Electronic Engineering, National University of Ireland Maynooth,
Maynooth, Co. Kildare, Ireland.
email: niall.cahill@eeng.nuim.ie, rlawlor@eeng.nuim.ie

ABSTRACT

In this paper a novel approach to doubletalk detection (DTD) is presented. This approach uses a modified Non-Negative Matrix Factorization (NMF) technique originally developed for monaural sound source separation to perform DTD. The efficacy of this approach is demonstrated through experiments using real room impulse responses (RIRs). The properties of this algorithm are then discussed with reference to experimental results.

Index Terms— Non-Negative Matrix Factorization (NMF), Doubletalk Detection (DTD), Acoustic Echo Cancellation (AEC).

1. INTRODUCTION

In telecommunications acoustic echo occurs when speech from one far end participant is broadcast into an enclosure at the opposite or near end user and is picked up by the near end microphone. This echo signal is then transmitted back to the far end user. If the all round time delay is large or the loudspeaker microphone coupling is high this returning echo causes significant annoyance for the far end user. Acoustic echo is particularly problematic for hands free telephony.

Most digital signal processing approaches to Acoustic Echo Cancellation (AEC) model the enclosure using a FIR filter with time varying coefficients to match and then cancel the echo signal. The coefficients are usually updated using an adaptive algorithm, which updates by comparing the incoming reference signal from the far end user with the echo signal picked up by the near end microphone [1].

During stable enclosure conditions and no doubletalk the above approach is sufficient for good echo cancellation. If however there is doubletalk, concurrent near end speech and echo signal, the adaptive filter will diverge away from the optimal echo canceling coefficients. To prevent this divergence Doubletalk Detectors (DTD) are used in conjunction with adaptive filters to pause adaptation and fix the filter coefficients during periods of doubletalk [1].

Common approaches to DTD include energy calculations with thresholding [2] and cross correlation [3]. Cross correlation techniques exploit the fact that the far end and near end speech are approximately uncorrelated. It also uses the fact that if the adaptive filter has converged the reference and residual error should be uncorrelated. Therefore any correlation detected between the reference signal and the residual error signal above a certain threshold will signify that further adaptation is needed. A normalized version of the correlation algorithm is presented in [4] with a focus on explicitly detecting doubletalk regions.

Echo path changes cause problems for DTD. Many of these techniques erroneously interpret echo path changes as doubletalk [1]. Pausing the filter adaptation during these periods will stop the algorithm converging to the new optimal filter coefficients. This will allow much echo to return to the far end user. Clearly this is undesired.

We present a novel technique for DTD that is completely immune from echo path change/doubletalk ambiguity. It is based on an alternative approach to AEC first presented in [5]. In [5] it was shown that building a time-varying NMF basis of the reference far end signal magnitude spectra, merging this basis with a static near speaker basis and then using NMF with this merged basis on the microphone signal spectrum can produce echo suppression. We show here that using this same approach both AEC and DTD can be realized. We also present a detailed description of the AEC algorithm first mentioned in [5].

The rest of this paper is organized as follows, in section s 2 and 3 we outline our algorithm in detail, in section 4 we present the doubletalk detector, in section 5 we describe the experiments undertaken and in sections 6 and 7 we discuss this work and conclude.

2. NMF AND MONAURAL SOUND SOURCE SEPARATION

Non-Negative matrix factorization is an approach for decomposing multidimensional non-negative data [6]. It works by approximating a data set $V \in \mathbb{R}^{\geq 0, M \times N}$ as a multiplication of two matrices $W \in \mathbb{R}^{\geq 0, M \times R}$ and $H \in \mathbb{R}^{\geq 0, R \times N}$.

$$V \approx W \cdot H. \tag{1}$$

The rank of the approximation can be reduced or increased by varying R; the number of columns in W and rows in H. This usually decreases or increases the reconstruction error depending on the data set. The decomposition is unique in that it enforces a non-negative constraint on W and H. This results in a parts based decomposition where the parts sum to form the whole [6].

In [7] a multiplicative gradient descent method for achieving NMF was presented which took advantage of the inherent non-negativity of the product of two non-negative values to impose a non-negative constraint on the data. One of the cost functions used in [7], is a generalized version of the Kullback-Leibler divergence,

$$D(V\|W,H) = \left\| V \odot \log\left(\frac{V}{W \cdot H}\right) - V + W \cdot H \right\|_{Fro}, \quad (2)$$

where $\odot$ is the Hadamard product. It can be minimised using the following multiplicative update rules for H and W, derived in [7],

$$H = H \odot \frac{W^T \cdot \left[\frac{V}{WH}\right]}{W^T \cdot 1}, \quad W = W \odot \frac{\left[\frac{V}{WH}\right] \cdot H^T}{1 \cdot H^T} \quad (3)$$

These update rules are iterated until a user-defined number of iterations have been reached. The number of iterations is usually picked to occur when the value of cost function D reaches a low value. H and W are updated alternately, as their objective functions are convex separately but not together. An advantage of these multiplicative updates is that no update step tuning is required.

The matrices H and W will individually express different aspects of the factorization. The columns of W will contain the basis for the data and the rows of H will contain the activation pattern for each basis or the contribution of each basis to the data over time. When multiplied the data is reconstructed with a small error (depending on R and the data).

There are many applications of NMF in the literature here we focus on one specific application which is Monaural Sound Source Separation (SSS). The goal of Monaural SSS is to separate out single sources of sound from one mixture. One supervised approach to achieve source separation is to train models of the speakers in a mixture a priori and then use these models to match the contribution of each speaker in the overall mixture [9]. In this framework and with the non-negative nature of the audio spectrogram, NMF has been used to train the speaker specific time–frequency domain models and to match the models of each speaker to their contributions in mixtures [8][9].

Using this approach separation is achieved in two stages; first, separate low rank W matrix bases are trained for each individual speaker. This is done by acquiring a sequence of spoken speech from each speaker, calculating a spectrogram for each sequence and performing NMF decomposition on each spectrogram separately. The resultant W matrices (one for each speaker) are then concatenated into a large W matrix called W_{train}. The second stage is the separation stage or matching stage where a mixture of speech, containing known speakers, is separated into individual sources. This is achieved by performing a NMF decomposition on the speech mixture using W_{train} from the training stage. Throughout this factorization W_{train} is fixed with only the H matrix updated. This process causes the basis of each speaker to match the mixture spectral energy corresponding to the contribution that speaker made to the mixture.

After a prescribed number of iterations have been reached, W_{train} is separated back to the individual W matrices of the speakers and then multiplied by the corresponding portion of the H matrix from the separation stage. The resultant V matrices are combined with the original phases of the mixture and resynthesised leading to renditions of the original sources.

3. AEC USING NMF

In conventional approaches to Acoustic Echo Cancellation the problem is modeled in the following way. The near end microphone signal m is comprised of the echo or desired signal d and the near end speech signal v,

$$m(n) = d(n) + v(n), \quad (4)$$

where n is the time index. The desired signal or echo d is modeled as a convolution of the reference signal x with a room impulse response (RIR) $\mathbf{h}$. This can be expressed as follows,

$$d = \mathbf{h}^{\mathrm{T}} x, \quad (5)$$

where,

$$\mathbf{h}^{\mathrm{T}} = [h_0, h_1, \ldots, h_{L-1}],$$
$$x = [x(n), x(n-1), \ldots, x(n-L+1)],$$

and L is the echo path length. The error e between the estimated filter $\mathbf{w}$ and $\mathbf{h}$ is defined as,

$$e(n) = m(n) - \mathbf{w}^{\mathrm{T}} \boldsymbol{x}, \quad (6)$$

this is used to update the L coefficients of $\mathbf{w}$ for LMS based algorithms. The estimated echo signal can then be subtracted from the loudspeaker echo thus canceling it. For this submission we neglect the affects of non-linearity's and noise both measurement and local.

In [5] a novel technique for AEC was presented which is based on a different approach than the adaptive filter style approaches prevalent in the literature. NMF trained bases

for the reference and nearend speech are used to match and remove echo in the magnitude time-frequency domain. This pattern-matching capability of NMF has been used to extract sources from monaural mixtures of known speakers as described in section 2. In [5] it was shown that a NMF basis model trained with anechoic speech can be used to match echoic versions of the same speech. This fact was exploited for AEC by training a time varying spectral magnitude NMF basis of the reference or far end signal x and using this basis to match and remove echo spectral energy in the near end microphone signal m.

This algorithm has two stages namely training and matching similar to the monaural SSS scenario. Before training the NMF bases the incoming far end signal x is partitioned into overlapped contiguous frames and each frame is transformed into the magnitude time frequency domain,

$$X(f,k)=\left|\sum_{t=0}^{T} x(st)w(n-t)e^{\frac{-2\pi jkt}{T}}\right|, \quad (7)$$

where X is the short-time Fourier transforms of x, f is the frequency bin index, k is the frame index and T is the framesize. The stepsize is controlled by s with the window index t. As more reference signal arrives at the near end user from the far end some previous frames can be buffered and concatenated to form a larger data matrix $X \in \mathbb{R}^{\geq 0, F\times K}$. K and F denote the number of frames in X and the number of frequency bins respectfully. From this a NMF decomposition is performed on X which will yield the far end speaker or echo basis B_x,

$$X \approx B_x G, \quad (8)$$

A new B_x is created for each new frame of far end speech with the F-1 previous frames. The G matrix is discarded each time. Each new B_x is combined with a second basis B_n for the near end speaker. This basis (B_n) will be trained a priori on independent speech utterances and remain static throughout the AEC process. The purpose of th B_n basis is to match any near end speech during doubletalk. R_2 denotes the rank of this matrix. We represent the combined training basis for use on the near end microphone signal as B_m.

$$B_m = [B_x \; B_n], \quad (9)$$

From here the matching or echo removal stage begins. The near end microphone signal m is partioned into contiguous frames with overlap and transformed into $M(f,k)$ using equation (7) .Then using NMF with a fixed B_m each $M(f,k)$ is factorized generating G_m,

$$M(f,k) \approx B_m G_m, \quad (10)$$

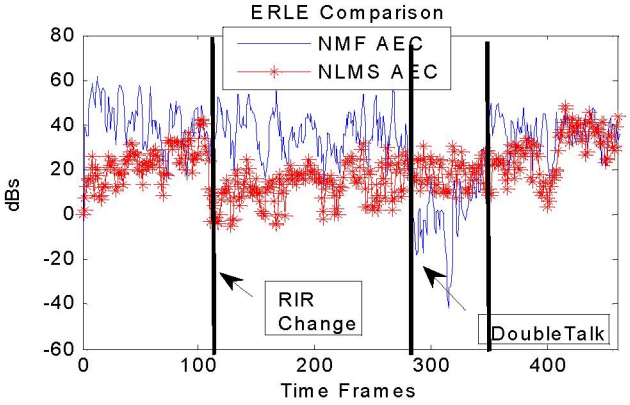

Figure 1: ERLE comparison between NMF AEC and NLMS AEC during doubletalk and echo path change.

where G_m is the gain matrix for B_m. After this process each G_m is separated into G_x and G_n and B_m. is separated back into B_x and B_n,

$$B_m^{F\times(R+R_2)} = [B_x^{F\times R} B_n^{F\times R_2}], \quad (11)$$

$$G_m^{(R+R_2)\times 1} = [G_x^{1\times R} G_n^{1\times R_2}]^T, \quad (12)$$

the magnitude spectrum of frame k of the output signal Y is then computed as:

$$Y(f,k) = B_n G_n, \quad (13)$$

each frame of the output signal y then resythesised using the IFFT with the phases calculated during $M(f,k)$ and a simple overlap and add scheme.

For each frame of speech returned to the far end user two NMF decompositions are performed (see section 2), one during the training of the echo basis and a second during the echo matching stage. These decompositions differ in there execution. During the training stage, each new B_x, and G are iterated alternately until a number of iterations have been completed then G is discarded. At the start of the matching stage however B_m is fixed and only the G_m updates are iterated. When the G_m iterations have been completed we let B_m be updated once or twice. We found these late iterations of the B_m update to be the key in achieving good echo cancellation using NMF. This is because the residual spectral magnitude energy not captured due to a fixed B_m at the end of the G_m iterations is subsequently captured by B_m by the late iterations of its update. Moreover the residual energy captured is primarily located in the B_x portion of B_m . We believe this is because after the G_m iterations the B_x portion of the basis is now more easily able to match the remaining residual echo energy. Furthermore B_n will generally be converged to zero making it difficult for the multiplicative updates to increase the energy in this region of the basis.

In Figure 1 a comparison between our NMF AEC algorithm and a conventional NLMS adaptive filter

algorithm in terms of Echo return loss ERLE is presented. ERLE as defined below,

$$ERLE = 10\log_{10}\left(\frac{E\{y^2(n)\}}{E\{e^2(n)\}}\right), \qquad (14)$$

The main advantages of our system are revealed through Figure 1. During echo path change the ERLE value for the NMF stays steady whilst the NLMS decreases as it reconverges to the new RIR. Also our approach does not need an initial lead in time to start removing the echo. These benefits are derived from the fact our algorithm at no stage estimates the room impulse response, it matches echo speech using a basis trained on the incoming reference far end speech. A disadvantage to our approach however is during doubletalk. Here the NLMS outperforms NMF. This is because if the NLMS has converged to a good estimate of the RIR before doubletalk and a good doubletalk detector is in place the estimated echo will cleanly remove the echo. The NMF approach however has to separate out the speech from the echo using a pre-trained independent speaker basis and an echo basis. This will lead to some crosstalk between the two respective bases. The lower ERLE value is due to a combination of decreased echo removal and echo removed by the near end basis.

4. DOUBLETALK DETECTION USING NMF

The AEC approach outlined in section 2 uses a second basis B_n to match any near end speaker speech to prevent it being removed by the echo basis B_x. In theory when the near user speaks this basis will be used to capture the entire near speech. This is because B_n will be closer in representation to the near end speech than the echo signal basis. It is straightforward therefore to assume that during periods of doubletalk the total energy E of each output frame of B_n will be high relative to periods where the near end speaker is inactive. This is illustrated in Figure 2, here the total magnitude of the clean nearend utterance is compared, in a frame wise manner, with the output of the NMF AEC approach. From this graph it is clear during the doubletalk period the envelopes of the two signals are highly correlated and have relatively the same magnitude.

To exploit this property for doubletalk detection we propose using a energy threshold to determine if near end speech is present in a frame or not,

$$E(k) = \sum_{f=0}^{F} Y(f,k)^2, \qquad (15)$$

$$DTD(k) = \begin{cases} 1, & \mathrm{E}(k) > \gamma, \\ 0, & \text{otherwise}, \end{cases} \qquad (16)$$

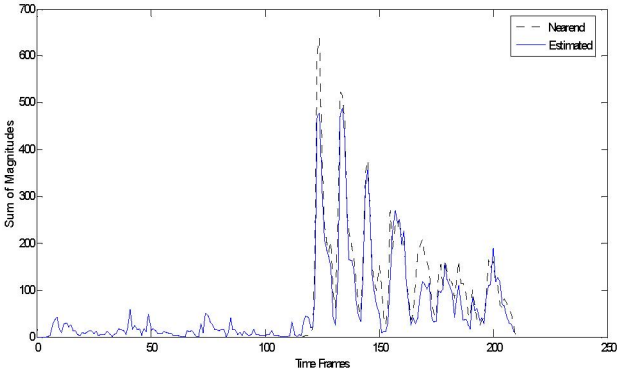

Figure 2: Comparison between NMF AEC total frame magnitude output and the clean near end speech signal AEC total frame magnitude. The data shows the sum of the magnitudes of each frequency for each frame over time

where DTD is the indicator function and γ is the threshold. If the energy E of a particular output frame k is greater than γ then DTD is set to one. This will signify that this frame is considered to contain near end speech. If E is below the threshold the frame is considered echo only. This detector also is dependent on the parameters set for the AEC NMF algorithm. Discussion of the performance of this algorithm will be provided in later sections.

5. EXPERIMENTS

The goal of the experiments is to analyze the performance of the novel doubletalk detector (DT). Real room impulse responses from the Mardy [10] database and speech from the TIMIT database [11] were used to create 3 test mixtures. Each mixture contained a background echo, which spanned the entire timeline of the mixture; see Figure 3a). Doubletalk regions were inserted into each mixture to test the performance of the DT. This required the introduction of a separate speaker convolved with a different RIR into the mixture. A change in the impulse response was imposed into each mixture at 42000 samples. This was forced on the far end echo signal by a sudden change in the RIR used to filter the far end speech.

The parameters of the algorithm were as follows: F was set to 4, B_x 20, B_n 2, number of iterations set to 40 and the B_m matrix is allowed to update twice at the end of each 40 G_m updates. The window size was 64 ms or 1024 samples for 16 kHz sampling rate with a 50% overlap between frames. B_n was trained using speech independent from the speech used to create the mixtures. The threshold of the DT γ was set to 5.

The algorithm and all work presented was implemented using Matlab. The results for the doubletalk detection task is displayed in Figure 2 b), c), d), where the regions of doubletalk detected using our approach regions are superimposed on the clean near end signal without the echo which is removed for clarity.

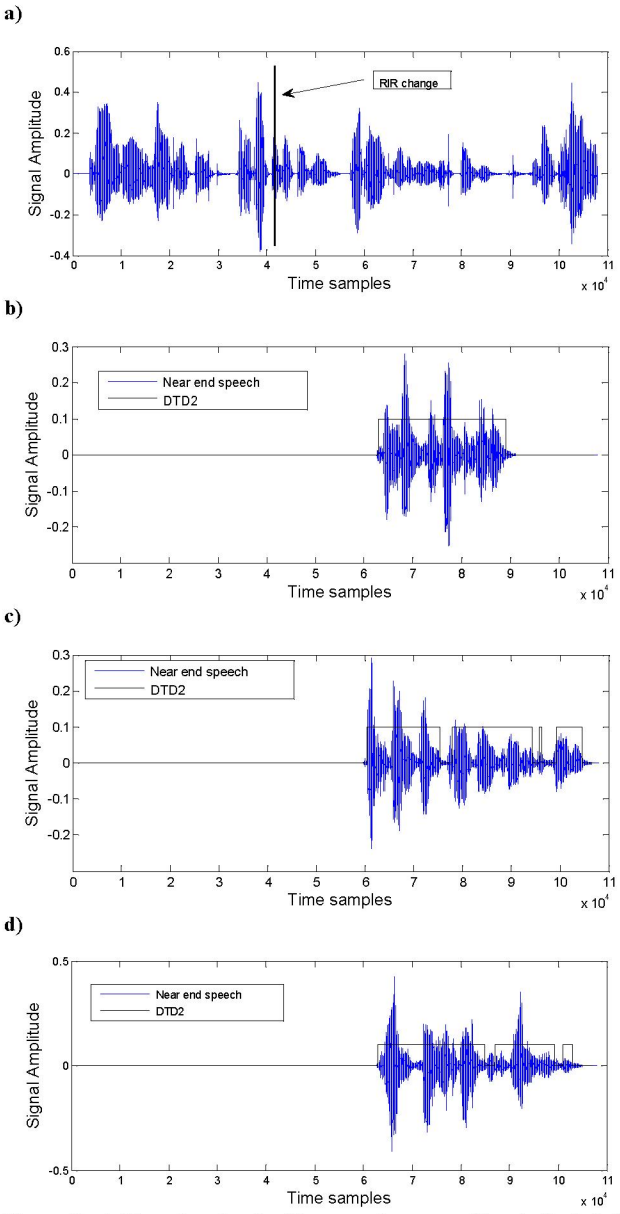

Figure 2: a) The echo signal with room change position indicated. b), c), d) near end talker utterance with DTD superimposed.

6. DISCUSSION

It can be seen in Figure 2 b), c) and d) that the doubletalk detector correctly labeled the vast majority of the doubletalk frames correctly. Moreover it is clear from Figures b), c), d) that during echo path change no false positives were generated, as expected. Also our approach does not need an initial lead in time to start detecting doubletalk correctly.

Changing the value of γ will increase the sensitivity of the DT and may cause some false positives to appear therefore it is important to set this parameter carefully. Further work will involve comparing this algorithm rigorously with other approaches using the objective approach first outlined in [12].

To provide full AEC this algorithm would be coupled with a conventional adaptive filter approach. This approach would enable adaptation to pause during doubletalk. Furthermore this hybrid system could mitigate some of the distortion introduced by the NMF AEC approach outlined in section 3 by combining it with a NLMS style approach.

Computational load is a major issue for AEC systems in low latency communications system. The DT approach presented here is computationally intensive. Further work is needed to quantify the load and reduce it.

7. CONCLUSIONS

In this paper a novel Doubletalk Detector was presented. Results show that this approach can detect doubletalk accurately and is immune to echo path changes. Experimental mixtures using real RIRs demonstrated the effectiveness of this approach.

8. REFERENCES

[1] S. Haykin and B. Widrow, *Least-mean-square adaptive filters*, Wiley-Interscience, Hoboken, N.J., 2003.

[2] D. L. Duttweiller, "A twelve-channel digital echo canceler," IEEE Trans. Comm., vol. 26, pp. 647-653, May 1978.

[3] H. Ye, and B. X. Wu, "A new double-talk detection algorithm based on orthogonality theorem," IEEE Trans. Comm., vol. 39,pp. 1542-1545, November 1991.

[4] J. Benesty, D. R. Morgan, and J. H. Cho, "A new class of doubletalk detectors based on cross-correlation," IEEE *Trans. Speech Audio Processing*, vol. 8, pp. 168-172, March 2000.

[5] N. Cahill and R. Lawlor, "A novel approach to acoustic echo cancellation" In *Proc. 16th European Signal Processing Conf. (EUSIPCO-08)*, Lausanne, Switzerland, September 2008. (to appear)

[6] D.D. Lee and H.S. Seung. "Learning the Parts of Objects by Nonnegative Matrix Factorization", in *Nature* 1999 (401):788.

[7] D.D. Lee and H.S. Seung, "Algorithms for non-negative matrix factorization", in *Advances in Neural Information Processing Systems* 13, 2000.

[8] P. Smaragdis, "Convolutive Speech Bases and their Application to Supervised Speech Separation", *IEEE Trans. on Audio, Speech and Language Processing*, Vol. 15, Issue 1, pp. 1-12, January 2007

[9] P. Smaragdis, "Discovering auditory objects through non-negativity constraints," in SAPA, 2004.

[10] J. Wen, N. D. Gaubitch, E. Habets, T. Myatt and P. A. Naylor. Evaluation of Speech Dereverberation Algorithms Using the MARDY Database. In *Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC-06)*, Paris, France, September 2006.

[11] W.M Fisher, G.R. Doddington, and K. Goudie-Marshall, "The DARPA Speech Recognition Research Database: Specifications and Status," *Proceedings of DARPA Workshop on Speech Recognition*, pp. 93-99, Feb. 1986.

[12] J. H. Cho, D. R. Morgan, and J. Benesty, "An objective technique for evaluating doubletalk detectors in acoustic echo cancelers," *IEEE Trans. Speech Audio Processing*, vol. 7, pp. 718-724, Nov. 1999