16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

A NOVEL APPROACH TO ACOUSTIC ECHO CANCELLATION

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 single microphone Acoustic Echo cancellation (AEC) is presented. This approach performs AEC by employing techniques developed for monaural sound source separation. It is shown that the AEC problem can be cast in a monaural sound source separation framework and through this framework significant echo suppression can be achieved. The new approach is evaluated through experiments on simulated data.

1. INTRODUCTION

With the proliferation of hands free mobile communications and VoIP the issue of Acoustic echo cancellation has become an increasingly important topic for both industry and academia. Acoustic echo occurs in full duplex communication when speech from a far end participant *x(t)* is broadcast into an enclosure at an opposite or near end user, is picked up by the near end microphone and retransmitted back to the far end user. The echo *y(t)* transmitted back to the far end user is dependent on the transfer function from loudspeaker to microphone through the enclosure. For long impulse responses fluid communication can become very difficult [1].

The loudspeaker-enclosure-microphone coupling (LEM) can be modelled as a time invariant linear FIR filter *h(t)*. However, it is known that small changes in the enclosure environment, such as the opening of a door, greatly affect the LEM filter. This likely possibility necessitates the use of an adaptive LEM filter to model the echo path over time. The echo signal *y(t)* can be stated mathematically as follows,

$$y(t) = n(t) + v(t) + \sum_{m=0}^{N-1} h(m)x(t-m), \qquad (1)$$

where N is the length of the impulse response, t is the time index for the output, *n(t)* is a noise term and *v(t)* is the near end speaker signal.

At present most AEC techniques use Least Mean Squares LMS and its many variants particularly normalised LMS (NLMS) [1] to estimate and update an estimate of the LEM filter coefficients. In general this is performed in a noise cancellation feedback structure whereby an estimate of the acoustic echo is estimated from the incoming reference speech and the input to the microphone from the enclosure. This estimate is then subtracted from the data before sending to the far end user.

There are a number of open problems with this approach [1]:

- For LEM filters with long impulse responses long estimation filters are needed, which can lead to convergence issues and large computational load.
- Noise in the reference signal and background noise from the near end can cause convergence problems for the adaptive algorithm.
- Changes in the LEM filter lead to periods where the adaptive algorithm must converge to new optimal parameters for the estimated LEM filter. This leads to a period of misadjustment, where a sub-optimal filter is used to remove the echo.
- When the near end user is speaking while the far end user is speaking the adaptive algorithm diverges away from suitable FIR coefficients. This is known as doubletalk (DT) in the literature.

A number of techniques have been developed to obviate or control these problems [1][2]. In general these techniques introduce trade offs into the overall AEC system.

Presented here is an alternative approach to AEC. A monaural sound source separation (SSS) technique based on non-negative matrix factorisation (NMF) is adapted to perform AEC. It is shown that this approach can lead to significant echo reduction.

This paper is organised as follows. In section 2 monaural sound source separation is described followed by NMF in section 3. In section 4 AEC using monaural SSS and NMF is explored followed by experiments and discussion in sections 4 and 5.

2. MONAURAL SOUND SOURCE SEPARATION

The goal of monaural or one-microphone sound source separation is to completely separate an arbitrary number of sound sources using only one mixture of the sources. The constraint of one mixture makes this task very challenging. Using only one mixture prohibits the use of any spatial information and prevents the application of well-established multi-sensor blind source separation techniques such as Independent component analysis (ICA). Undetermined (less sensors than sources) blind source separation BSS techniques have been developed based on sparsity and spatial cues. These techniques also require at least two

mixture signals. Spatial cues are central to all multi-sensor techniques because the sources are assumed to have independent spatial signatures. These spatial cues are then used to invert a mixing matrix, as in ICA, or to group components in the sparsity-based techniques.

One emerging theme for monaural SSS, is the use of prior information about the source signals to perform separation. This deviates from the ideal of blind sound source separation but is considered necessary in light of the constraints one-mixture imposes. One general framework has been to train bases or models on training data for each speaker a priori and then match these models with a mixture containing these speakers [3-7].

Within this framework, many different approaches to modelling and grouping/matching of mixture components have been attempted. Many techniques are spectrogram based and model the mixture speech as individual speakers basis multiplied by a time varying gain. A number of researchers have used non-negative matrix factorisation (NMF) [3] or sparse NMF (SNMF) [4] to build up speaker independent bases and then these bases are used to decompose mixtures in the time-frequency domain. Other researchers have trained Markov models for individual speakers and used these models to update time varying subband gains to separate sources [5]. Another approach is to use convolutive NMF [6] bases, which extend the time extent of ordinary NMF bases, for training and matching. Time domain bases have also been used. In [7] time domain basis were trained using ICA and then matched to the mixture using maximum likelihood.

For the work presented here we investigate the application of such an approach to the AEC problem. NMF is utilised to perform both training and matching in the audio spectrogram similar to as described above.

3. NON NEGATIVE MATRIX FACTORISATION

Non-Negative Matrix Factorization (NMF) is a linear data analysis technique for non-negative data [8]. The non-negativity constraint of this factorization results in a parts based/additive decomposition of the data where the individual decomposed parts sum together to form the original data. This decomposition provides a more intuitive representation of the underlying data [8]. 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 $H \in \mathbb{R}^{\geq 0, R \times N}$.

$$V \approx W \cdot H. \tag{2}$$

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 process of estimating W and H is an optimization problem. Lee and Seung [9] introduced two approaches for estimating W and H each based on a separate cost function. The Euclidean distance between V and WH was one of these cost functions and the second, which was used throughout this work, is a generalized version of the Kullback-Leibler divergence,

$$D(V \| W, H) = \left\| V \odot \log\left(\frac{V}{W \cdot H} \right) - V + W \cdot H \right\|_{Fro}, \tag{3}$$

where $\odot$ is the Hadamard product. The goal of the optimization is to minimize this cost function with respect to W and H whilst imposing the non-negativity constraint. From equation (3) the following multiplicative update rules were derived in [9] to calculate H and W,

$$H = H \odot \frac{W^T \cdot \left[\frac{V}{WH} \right]}{W^T \cdot 1}, \quad W = W \odot \frac{\left[\frac{V}{WH} \right] \cdot H^T}{1 \cdot H^T} \tag{4}$$

These update rules are iterated until a prescribed number of iterations has been reached. The updates are alternated between H and W, as the objective functions for each are convex separately but not together. Because of the multiplicative updates no update step tuning is needed. The number of iterations specified is data/user dependent and usually picked to occur when cost function D reaches a user-defined threshold.

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).

Monaural SSS can be performed using NMF 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 a 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 leads to the basis corresponding to each individual speaker to mainly characterize the mixture spectral energy corresponding to the contribution, which 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.

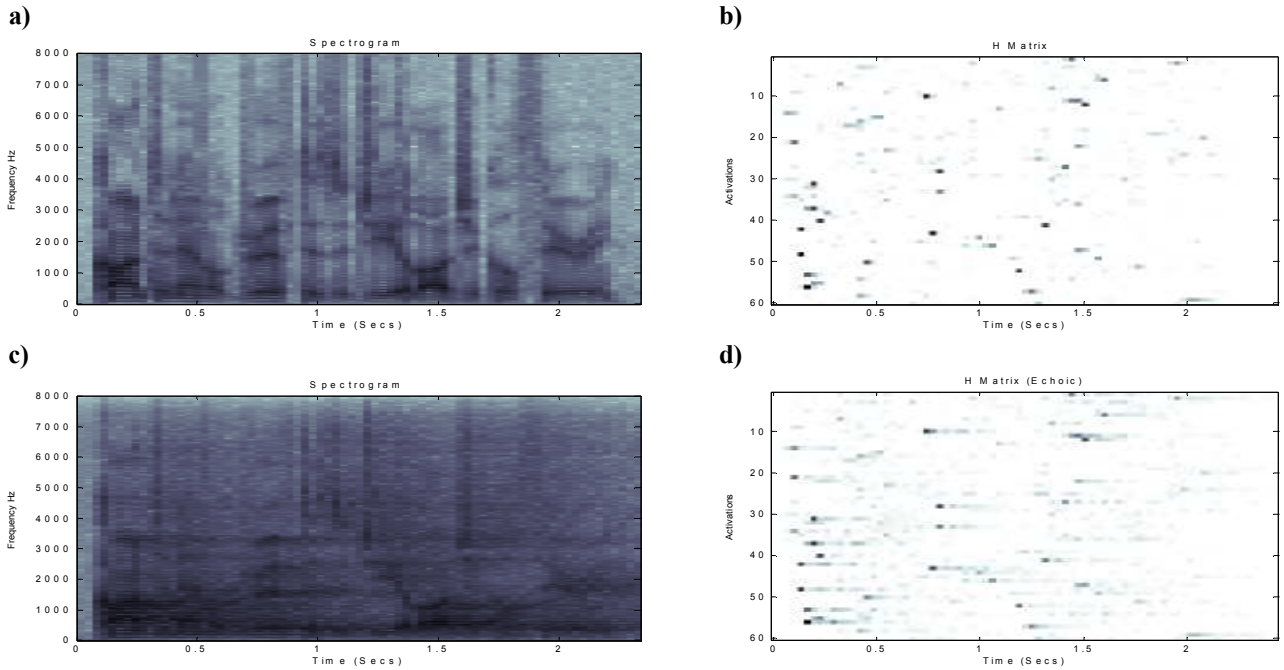

Figure 1: a) Spectrogram of speech recorded in an anechoic environment. b) H matrix from NMF performed on spectrogram in a), c) Spectrogram of speech in a) convolved with a room impulse response d) H matrix for NMF performed on spectrogram in c) using W matrix from b).

The best separation performance using this approach is achieved when the sources in the mixture are spectrally dissimilar [6]. An example would be a two-source mixture, which contained one male and one female speaker. These spectrograms have a greater level of dissimilarity than say between the spectra of two males or two females due to the different pitch tracks and formants etc. As a result of this the trained *W* matrix bases for the male and female speech are more easily able to distinguish and better represent their respective contributions in the mixture. This issue of spectral dissimilarity was shown to be an important factor affecting the performance of this monaural SSS algorithm [6].

4. AEC AND MONAURAL SOUND SOURCE SEPARATION

The AEC problem is a special case of the monaural sound source separation problem. In AEC the goal is to remove the echo from the speech transmitted back to the far end speaker with or without doubletalk. This can be taught of as a monaural SSS problem with 2 sources; the echo and the local speaker. Following on from this a basis for the echo signal can be trained using the incoming reference speech and a separate basis can be trained for the local speaker signal using pre-recorded speech data. Together these bases can be used to separate out or remove the echo from the returning microphone mixture. An advantage of using this approach for AEC is the fact that the reference basis will be trained using the actual speech being used to excite the LEM. This will facilitate better matching of the echo spectrogram and thus removal. This will alleviate the problem of spectral dissimilarity described in section 3.

Another aspect of this approach is the effect reverberation has on the *H* matrix from NMF decompositions of audio spectrograms. The rows of *H* contain a time varying gain for each basis in *W*. These varying gains contain the contribution each basis makes to the mixture spectrogram over time. For anechoic speech the *H* matrix is usually a sparse matrix with activations usually occurring in single spikes over time. However if the same *W* matrix was used for an echoic version of the spectrogram the activations in *H* become smeared. Figure 1 illustrates this effect. This is because the echoes in the speech manifest as repeated/smeared copies of the anechoic spectrogram. The NMF represents these echoes as repeated and scaled copies of the original *W* basis over time. This property of the NMF audio spectrogram enables the basis to be trained on anechoic speech and then can be used to separate echoic speech. This applies to AEC as the reference signal first excites a LEM system before reaching the microphone.

Using this approach the effect of misadjustments/ enclosure changes will be mitigated. This is because NMF continuously adapts to the data present in the spectrogram and does not estimate the LEM filter; therefore it does not require further samples of the reference/microphone signal to converge to the new room response like LMS. This also means that the length of the impulse response is insignificant, as NMF will use the best available bases to match the contribution from long impulse responses i.e. the reference signal basis. LMS techniques usually fix the length of the estimation filters for the case of long LEM filters. In addition, using this approach Doubletalk will have less effect on this system, as there is a local speaker basis to match any near end speech.

a)

Fear end	Far end (Echo)	ERL dBs	NLMS ERLE dBs	NMF ERLE (dBs)
Female 1	Female 2	1.1725	33.0431	30.3651 (see figure 2)
Male 1	Male 2	1.6610	33.9400	34.0072
Female 3	Male 3	1.5097	34.3433	32.7138
Average		1.4477	33.7755	32.3620

b)

Fear end	ERL dBs	NLMS ERLE dBs	NMF ERLE dBs
Female 1	1.1725	22.5088	36.7518
Male 1	1.6610	26.2779	38.0236
Female 3	1.5097	24.5881	40.3750
Average	1.4477	24.4583	38.3835

Table 1: a) Results of experiments with Doubletalk mixtures. b) Results of experiments with simulated Room changes.

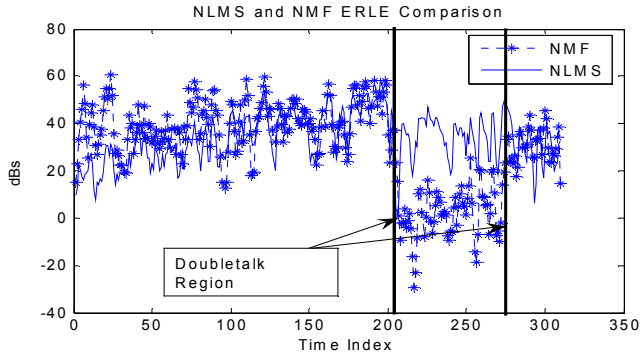

Figure 2: ERLE performance for NMF AEC and NLMS AEC with doubletalk. Mean ERLE is in Table 1 a)

5. ALGORITHM IMPLEMENTATION

The NMF based AEC approach described in the previous section was implemented in Matlab. The incoming reference speech *x(t)* (far end speech) was segmented into frames 64 ms long with a 50 % overlap between adjacent frames. A NMF decomposition is performed on the magnitude spectrum of this frame with the magnitude spectrum of previous buffered frames. The purpose of these extra, buffered frames is so all of the echo tail can be matched by this reference signal basis. Then the reference signal basis is merged with a near end speaker basis to form the complete *W* matrix and is then applied to the incoming near end microphone signal frame magnitude spectrum. After the *H* update iterations have been completed with a fixed *W* matrix, the speaker basis and its *H* component are multiplied and resynthesised using the phase of the microphone signal with the inverse Discrete Fourier transform and a simple overlap and add.

The number of reference signal basis vectors *R* was set to 40 with the number of speaker basis vectors set to two. The number of previous frames in each *V* was set to 6 with a new decomposition performed for each new incoming reference frame. The number of iterations of the *H* update for each new frame was set to 60. The output from this algorithm is an estimate of the near end speaker resynthesised from the speaker basis.

To increase spectrogram-matching performance the *W* matrix or basis is allowed to iterate twice at the end of the *H* updates. This improves the quality of echo signal matching greatly and leads to a significant increase in ERLE.

The speaker basis was trained a priori using sample sentences from different male and female speakers. The speech or speakers used to train the speaker basis were not then used in experimental mixtures. This algorithm was tested using simulated mixtures, which are described, in the next section.

6. EXPERIMENTS

The focus of the experiments in this paper is to demonstrate this approach for echo with and without far end speech and for changes in the LEM/RIR filter. We adopt the conventional AEC situation where the far end speaker speech is used to excite the LEM system at the near end user. For the experiments in this submission we neglect the effect of noise, both measurement and local background noise, on the overall system. We performed all processing in an offline/batch fashion.

Synthetic room impulse responses RIR were created using the mirror image method of creating room impulse responses [9]. The room was set up as a box room with dimensions 8m (length) × 7m (width) × 5m (height) and different frequency dependent absorption coefficients were set for each wall. A microphone was placed at (4, 3.5, 1.2) with a source emulating the near end loudspeaker placed 10 cm away (4.1, 3.5, 1.2) and another source representing the near end speaker placed 2.5 m away (6.5, 3.5, 1.2). This source would be used in experiments to simulate a change in the LEM filter. From this set-up three impulse responses (RT_{60}=120ms) from each source to the microphone were computed. Each impulse response was truncated to 1000 coefficients (60 ms with a 16 kHz sampling rate).

Mixtures were created using speech from different speakers; both male and female. These speakers were chosen arbitrarily from the TIMIT database [10]. Two AEC issues were examined, doubletalk and RIR changes. For doubletalk experiments each mixture had a near end speaker contribution and a main far end speaker (echo) contribution. Both these contributions were obtained by convolving separate sentences of speech with the respective RIRs. This is illustrated in Figure 4. To test LEM changes a sharp change in RIR was introduced in the echo mixture to simulate an enclosure change. For these experiments only far end speech was used.

For all experimental mixtures a comparison of our NMF AEC algorithm was made with a NLMS AEC system. This algorithm was configured with a 2000 tap filter length with the stepsize set to 0.5. During doubletalk the adaptation of the NLMS algorithm was stopped and the reference signal was filtered with the estimated RIR at that instant.

The results of the experiments were evaluated using objective energy ratios. To measure the input echo strength the Echo Return Loss (ERL) was calculated according to the following equation,

$$ERL = 10\log_{10}\left(\frac{E\{x^2(t)\}}{E\{y^2(t)\}}\right). \quad (5)$$

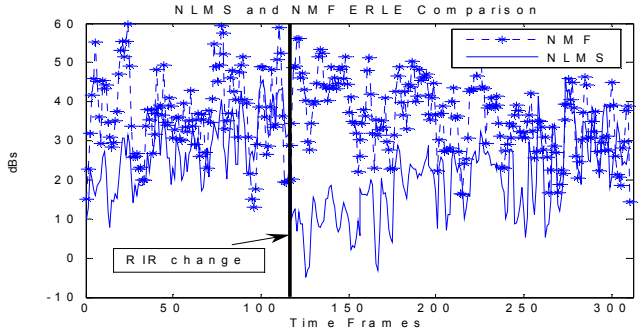

Figure 3: ERLE performance of NMF AEC and NLMS AEC for a change in room conditions

The performance of the algorithms was evaluated using the echo return loss enhancement measure ERLE. This ratio is a measure of the level of echo suppression and is defined as follows,

$$ERLE = 10\log_{10}\left(\frac{E\{y^2(t)\}}{E\{e^2(t)\}}\right), \qquad (6)$$

where *y(t)* is the echo signal and *e(t)* is the echo left after processing. The results of the experiments are tabulated in Table 1 a) and b). Plots of frame wise ERLE performance are given in Figure 3 and 4; with a plot of an example output from the two algorithms is given in Figure 4.

7. DISCUSSION

The results listed in Table 1a and 1b show that our NMF AEC approach has comparable performance to NLMS for doubletalk mixtures and superior performance during echo path changes.

In Figure 3 it is seen that a sudden change in the enclosure environment results in a sharp decrease in ERLE for NLMS whilst the NMF approach maintains its ERLE performance. This is because the NMF approach does not estimate the RIR and therefore, in the event of a echo path change continues to match echo spectral energy as before. This is also the case during the initial convergence of the NLMS were NMF has better performance.

The results in Table 1b of the Doubletalk experiments show NMF can provide echo cancellation. However as seen in Figure 2 ERLE falls for our approach. This is due to the high ratio of echo reference basis vectors to speaker basis vectors (for this submission 40:2). This caused some of the echo basis to capture some near end speech thus removing it from the output. As a result the error value increased leading to lower ERLE. The output however is largely free from echo but is distorted (see Figure 4) as some of its energy was captured by the reference/echo basis. Further work will involve improving this performance.

The NMF decomposition is relatively computationally intensive compared to NLMS style algorithms. More work is needed to reduce the computational load of this algorithm for implementation on real time processors. Other further work will involve finding the optimum value for the parameters of the algorithm such as *R*, the number of previous buffered frames and the number of iterations per frame.

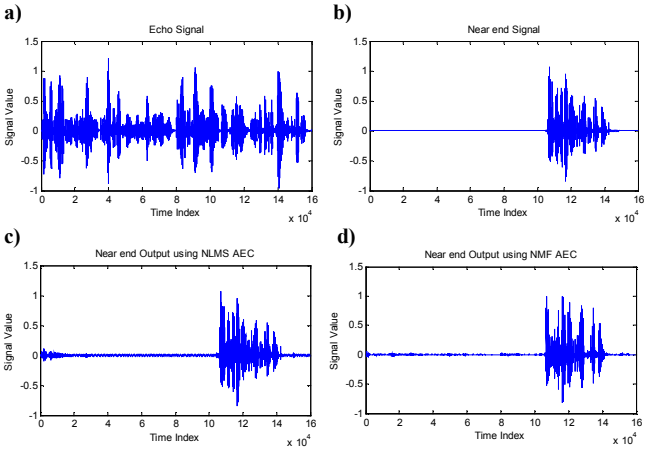

Figure 4: a) Echo signal (far end speech convolved with RIR), b) near end signal, male, c) Output from NLMS, d) Output from NMF (See table 1a mixture 1 for mean ERLE and Figure 2 for Frame based ERLE).

8. CONCLUSIONS

It is shown in this paper that the Acoustic Echo can be reduced using non-negative matrix factorisation in a monaural sound source separation framework. Results from experiments using synthetic data were used to demonstrate the performance of this approach.

REFERENCES

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

[2] P. Ahgren, "Acoustic echo cancellation and doubletalk detection using estimated loudspeaker impulse responses ". *Speech and Audio Processing, IEEE Trans.* **13**(6): 1231-1237. (2005).

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

[4] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *(INTERSPEECH)*, 2006.

[5] S. T. Roweis, "One microphone source separation," in *NIPS*, 2001, pp. 793–799.

[6] 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

[7] G. J. Jang and T. W. Lee, "A maximum likelihood approach to single channel source separation," *JMLR*, vol. 4, pp. 1365–1392, 2003.

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

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

[9] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, Apr.1979.

[10] 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.

[11] E. Vincent; R. Gribonval; C. Fevotte; "Performance Measurement in Blind Audio Source Separation", *IEEE trans on Speech and Audio processing.* Volume PP, Issue 99, 2005 Page(s): 1 – 8.