Blind Speech Separation in Convolutive Mixtures Using Negentropy Maximization

Vuong Hoang Nam, Nguyen Quoc Trung, Tran Hoai Linh
Hanoi University of Science and Technology
Email: namvh-fet@mail.hut.edu.vn

***Abstract*: This paper proposes a new method to address the problem of blind speech separation in convolutive mixtures in the time domain. The main idea is extract the innovation processes of speech sources by non-Gaussianity maximization and then artificially color them by re-coloration filters. Some simulation experiments of the 2x2 case are presented to illustrate the proposed approach.**

Keywords: Blind Signal Separation (BSS); Independent Component Analysis (ICA); FastICA; Negentropy Maximization.

I. INTRODUCTION

Blind source separation (BSS) is a technique to estimate original source signals using only sensor observations. If source signals are mutually independent and non-Gaussian, we can apply techniques of independent component analysis (ICA) to solve a BSS problem. Let us formulate the BSS model of convolutive mixtures. Suppose that N original sources are blindly mixed and observed at N sensors. We have the relations between the observations and the sources in time domain:

$$x_i(n)=\sum_{j=1}^{N}\sum_{k=0}^{\infty}h_{ij}(k)s_j(n-k)+\nu_i(n),\quad \forall i=\overline{1,N} \qquad (1)$$

where $x_i(n)$ is the observation at the ith sensor, $s_j(n)$ is the jth source and $\nu_i(n)$ is the additive noise. Denoting by $s(n)=\left[s_1(n),...,s_N(n)\right]^T$ the vector of orginal sources and by $x(n)=\left[x_1(n),...,x_N(n)\right]^T$ the observations at sensors, we have the convolutive BSS model in Z domain:

$$X(z)=H(z)S(z) \qquad (2)$$

where $X(z)$ and $S(z)$ are, the Z transforms of $x(n)$ and $s(n)$ respectively. The $N\times N$ matrix $H(z)$ consists of the transfer functions $H_{ij}(z)=Z\left[h_{ij}(n)\right]$ between the *j*th source and the *i*th sensor. In our model, we will assume no additive noise, all mixing filters $H_{ij}(z)$ are causal and FIR as well as the sources are stationary.

In convolutive BSS model, trying to extract the source signals is meaningless because the mixture (Eq. (2)) is not unique: An infinite set of couples $\left(H(z),X(z)\right)$ verifying the same assumptions yields the same output $X(z)$. Therefore, our aim is to estimate the contributions of all orginal sources in each sensor, e.g., $H_{ij}(z)S_j(z)$. Some author, [1-4], worked on the problem of convolutive BSS to deal with artificial colored signals and proposed a solution which consists in first estimating innovation processes by inverse filters, then building re-coloration filters to artificially color innovation processes in order to estimate contributions of each source signal in each sensor. In this paper, we apply this solution to a particular case: BSS for convolutive mixtures of speech. In our work, a more deeply analysis and study on this case has been made. The proposed model also deal with linear instantaneous mixtures by choosing zero as the order of all filters.

The remaining of this paper is organized as follow. In Section II, the detailed proposed approach is presented. The experimental results and discussion are

showed in Section III, while the conclusions are contained in Section IV.

II. THE PROPOSED APPROACH

We assume each speech source results from an innovation process colored by a speech production system modeled as a AR filter of order *P* [5-9]. Given a original speech source $s_j(n)$, we define its innovation process $e_j(n)$ as the error of the best prediction of $s_j(n)$, given from its past. The term "innovation" means that $e_j(n)$ contain all the new information about the process that can be obtained at time n. Then $s_j(n)$ is described as:

$$s_j(n) = \sum_{k=1}^{P} u_{jk} s_j(n-k) \quad + \; e_j(n) \tag{3}$$

or equivalently,

$$s_j(n) = \frac{1}{1 - \sum_{k=1}^{P} u_{jk} z^{-k}} e_j(n) \tag{4}$$

The relationship in *Z* domain:

$$S_j(z) = \left[U_j(z)\right]\left[E_j(z)\right] \tag{5}$$

where $E_j(z)$ is the Z transform of $e_j(n)$, $U_j(z)$ is a filter corresponding to the AR process.

In this paper, all mixing filters are supposed to be MA so that observed signals are outputs of ARMA processes, driven by innovation processes:

$$X(z) = \left[H(z)\right]\left[U(z)\right]\left[E(z)\right] \tag{6}$$

where $\left[E(z)\right] = \left[E_1(z),...,E_N(z)\right]^T$;

and $\left[U(z)\right]$ is a diagonal matrix defined as:

$$\left[U(z)\right] = diag\left(U_1(z),...,U_N(z)\right) \tag{7}$$

Furthermore, define $A(z) = \left[H(z)\right]\left[U(z)\right]$; we get:

$$X(z) = \left[A(z)\right]\left[E(z)\right] \tag{8}$$

Figure 1 depicts the system that produces the observed signals from the innovation processes. To simplify the notations, all filters $A_{ij}(z)$ in (8) are supposed to be MA because we can estimate a ARMA model based on the equivalent (long) MA [10]. We can see that there is no distinction between (Eq. (2)) and (Eq. (8)). Moreover, the innovations of speech sources are usually independent from each other as well as more non-Gaussian (super-Gaussian distribution) than original sources. Therefore, we can directly estimate innovation processes instead of speech sources by the non-Gaussianity maximization approach. The proposed approach consists of two main stage: innovation extraction stage and re-coloration stage.

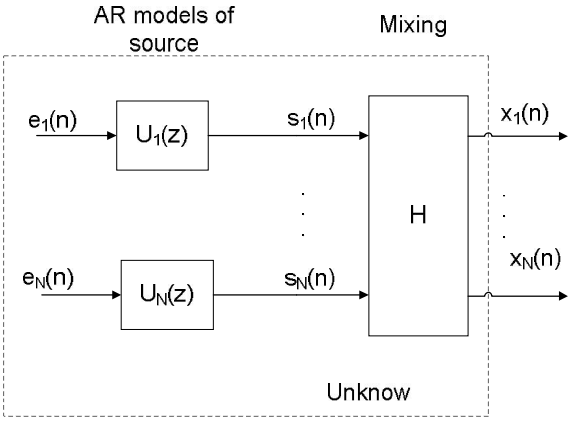

Fig.1. Schematic diagram of system producing observed speech signals from innovation processes

A. Innovation Extraction Stage

Each output signal $y(n)$ of the extraction stage is computed as following:

$$y(n) = \sum_{p=1}^{N} D_p(n) * x_p(n) = \sum_{p=1}^{N} \sum_{r=-R}^{R} D_p(r) x_p(n-r) \tag{9}$$

where $D_p(r)$, $p = 1,2,...,N$ are FIR inverse filters.

These filters are non-causal and MA in practice. In this stage, we use negentropy as the measure of non-Gaussianity [11-12]. If x is assumed to have zero mean and unit variance then the negentropy of x, denote by $J(x)$, can be approximated as following:

$$J(x) \approx \left[E\{G(x)\} - E\{G(\gamma)\}\right]^2 \quad (10)$$

where γ is a Gaussian variable of zero mean and unit variance, G is a suitable contrast function. The following choices of G have proved very useful [11-12]:

$$G_1(t) = \frac{1}{a_1}\log(\cosh a_1 t), 1 \le a_1 \le 2 \quad (11)$$

$$G_2(t) = -\exp\left\{-\frac{t^2}{2}\right\}, G_3(t) = \frac{t^4}{4} \quad (12)$$

By maximizing of the non-Gaussianity of the output signal $y(n)$, we can estimate an innovation process $e_l(n)$ of a speech source $s_l(n)$ up to a constant scale and delay under some conditions [1]:

$$y(n) = \alpha_l e_l(n - r_l) \quad (13)$$

For instantaneous mixtures, an algorithm named as FastICA, based on negentropy was proposed by Hyvarinen for blind source separation [11-12]. In [4], [20], authors extended this algorithm to convolutive mixtures by reformulating the problem using the instantaneous ICA model. At any time *n*, we define a column vector $\tilde{x}(n)$ by concatenating $(2R+1)$ time-delay versions of every observed signal:

$$\tilde{x}(n) = [x_1(n+R),...,x_1(n-R),..., x_N(n+R),...,x_N(n-R)]^T \quad (14)$$

which contain $M = (2R+1)N$ entries. We derive the M entry column vector $x'(n) = \left[x_1'(n),...,x_M'(n)\right]^T$ defined as:

$$x'(n) = B\tilde{x}(n) \quad (15)$$

where B is an whitening matrix chosen so that

$$E\left\{x_i'(n)x_j'(n)\right\} = \delta_{ij}, \forall i, j \in \{1,...,M\} \quad (16)$$

Eq. (15) may be considered as conventional whitening in FastICA. Using these definitions, the convolutive mixing model in (9) can be written:

$$y(n) = w^T.x'(n) = \sum_{m=1}^{M} w_m.x_m'(n) \quad (17)$$

where w is a $M-$ entry vector containing the coefficients of the FIR filters $D_p(r)$, $p = 1,2,...,N$ in a suitable order. Now we can estimate the convolutive model by applying the ordinary FastICA algorithm to the standard linear ICA model: Maximize the negentropy of $y(n)$ subject to $\|w\| = 1$

Re-coloration Stage

In this stage, we have to identify N non-causal re-coloration filters $C_k(z) = \sum_{r=-R'}^{R'} c_k(r)z^{-r}$, $k = 1,2,...,N$ and apply them to $y(n)$ in order to estimate contributions of $s_l(n)$ in each microphone. Thus, each source will have N contributions. The recovered signal of $s_l(n)$ is the most powerful contribution among its contributions. The contribution of $s_l(n)$ in the *k*th microphone is yielded by $C_k(z)*y(n)$. Let us denote by $d_k(n)$ the difference between the *k*th observation and the contribution of $s_l(n)$ in the *k*th microphone:

$$d_k(n) = x_k(n) - \sum_{r=-R'}^{R'} c_k(r)y(n-r) \quad (18)$$

Moreover, from (8), we have:

$$x_k(n) = \sum_{q=1}^{N}\sum_{r=0}^{L} a_{kq}(r)e_q(n-r) \quad (19)$$

where L is defined as the largest order of MA filters $A_{kq}(z)$ in (8).

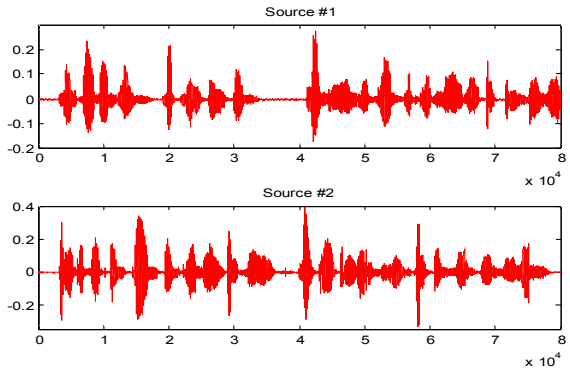

Fig 2. The speech source signals

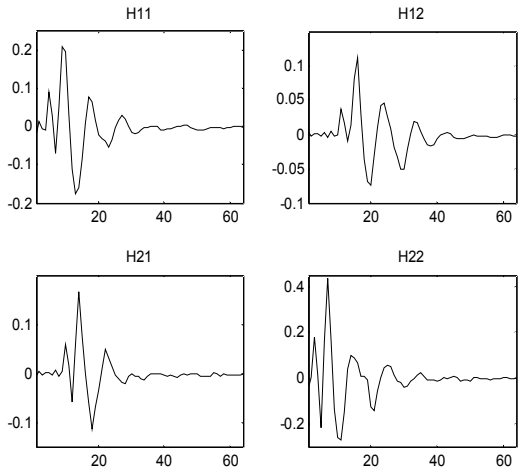

Fig.3. The four simulated mixing filters

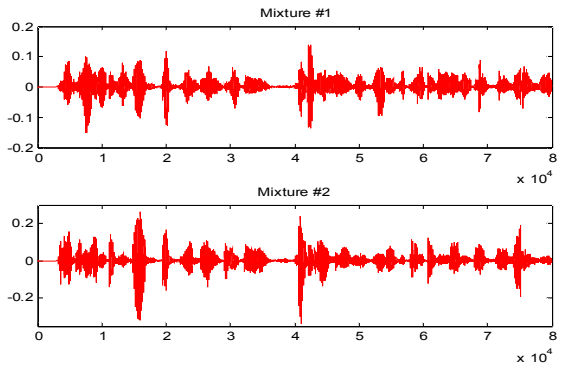

Fig 4 The mixtures of speech source signals

If R' is rather large so that $r_l \le R'$ and $R'+r_l \ge L$, combined with (13) and (19), Eq. (18) becomes

$$d_k(n) = \sum_{r=r_l-R'}^{r_l+R'} \left[a_{kl}(r) - \alpha_l c_k(r-r_l)\right] e_l(n-r) + \sum_{q\neq l}\sum_{r=0}^{L} a_{kq}(r) e_q(n-r) \quad (20)$$

The coefficients of the re-coloration filter $C_k(z)$ will satify the following condition:

$$a_{kl}(r) - \alpha_l c_k(r-r_l) = 0, \forall r \in [r_l - R', ..., r_l + R'] \quad (21)$$

The condition (21) is equal to the function $E\{d_k^2(n)\}$ is minimized. This can be done by a non-causal FIR Wiener-Kolmogorov filter that make the signal $y(n)$ be the closet to $x_k(n)$ in the mean-square sense. Therefore, we get:

$$c_k = R_{yy}^{-1} r_{yx} \quad (22)$$

where c_k is the recoloration filter coefficient vector, R_{yy}^{-1} is the autocorrelation matrix of the input signal $y(n)$ and r_{yx} is the cross-corelation vector of the input $y(n)$ and the desired signal $x_k(n)$.

B. The Deflation Procedure

In the proposed approach, we use a simple and efficient deflation procedure [1, 3, 4, 7, 8, 11-16]. After the successful extraction of the contributions of a source signal, we can apply the deflation procedure which removes the extracted signals from the mixtures. This procedure may be recursively applied to extract sequentially the rest of the mixing source signals.

C. The Overall Approach

From observations, the extraction stage yield a signal which only contain an innovation process up to

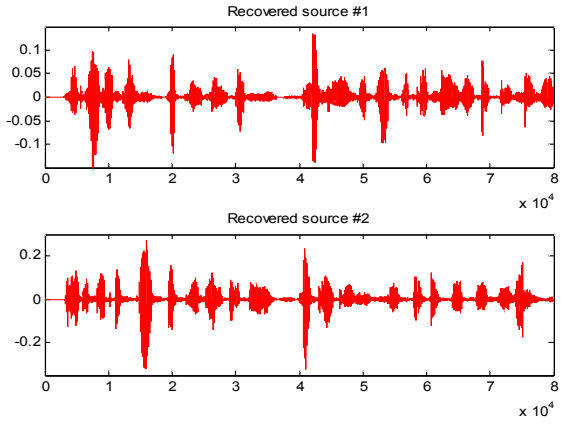

Fig 5. The estimates of speech source signals using G_3

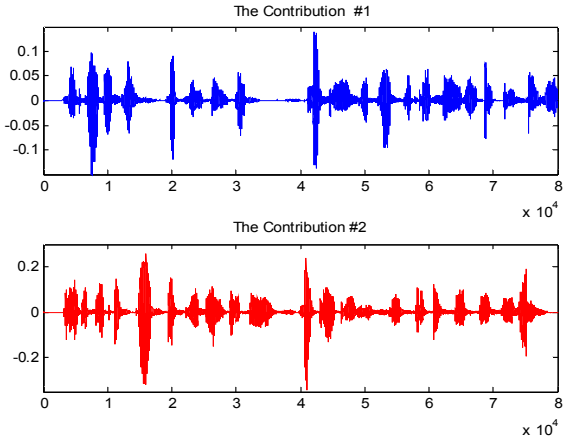

Fig 6. The true contributions of the speech signals

a constant scale and delay $y(n) = \alpha_l e_l(n - r_l)$.

The re-coloration stage is then applied to $y(n)$ and observations in order to estimate contributions of $s_l(n)$ in each microphone.

Remove the above contributions from the observations. Set $N = N - 1$. If $N > 1$ go back to Step 1, else quit.

D. Limitations of the approach

The above approach will be implemented to each signal frame. Ideally, it imposes the following conditions:

The signal frame size is larger than the order of mixing filters as well as that of speech production systems.

None of the system parameters change within a single frame.

(i) Mixing filters $H_{ij}(z)$ are minimum phase and matrix $H(z)$ has full rank [1, 14, 17]. This condition may be an unrealistic assumption.

In reality, the parameters of the speech production system $U_j(z)$ always change by tens of milliseconds while the order of the room acoustics $H_{ij}(z)$ may be equivalent to hundreds of milliseconds. When using a large frame size, it is impossible to equalize $y(n)$ with $\alpha_l e_l(n - r_l)$ because $U_j(z)$ varies within a single frame. Therefore, we have to use a frame size shorter than the order of realistic room acoustics, which enables us to equalize $y(n)$ with $\alpha_l e_l(n - r_l)$. Moreover, if the length of unmixing (inverse) filter is (very) long this method have a large computational load to compute as well as slow convergence speed.

Because of these limitations, this approach only yields good performance when mixing filters are not too long, so it is difficult to apply this approach in realistic acoustic environments.

III. RESULTS AND DISCUSSION

In our initial experiment, we created convolutive mixtures from two Vietnamese speech sources, as shown in Fig. 2, sampled at 16 KHz during 5 seconds. The simulated 64th-order mixing filters, [18], used in this experiment are depicted in Fig. 3. We used these responses to create the mixed signals as follows:

$$\begin{aligned} x_1(n) &= h_{11} * s_1(n) + h_{12} * s_2(n) \\ x_2(n) &= h_{21} * s_1(n) + h_{22} * s_2(n) \end{aligned} \quad (23)$$

The two mixed signals are shown in Fig. 4. To evaluate the performance of the proposed method, the Signal to Interference Ratio (SIR) is used. The

"Signal" is defined as the ideal (true) value $x_{ij}(n)$ of $x_{ij}^{'}(n)$ which is the estimated contribution of $s_j(n)$ in $x_i(n)$. The "Interference" is the deviation between $x_{ij}^{'}(n)$ and its ideal value $x_{ij}(n)$, e.g $x_{ij}^{'}(n) - x_{ij}(n)$. We define SIR(dB) of the estimation of a speech source signal $s_j(n)$ as follows:

$$SIR(s_j) = \max_i 10\log_{10}\left(\frac{E\{x_{ij}^2(n)\}}{E\left\{\left(x_{ij}^{'}(n) - x_{ij}(n)\right)^2\right\}}\right) \quad (24)$$

The length of the inverse filters as well as re-coloration filters should be chosen sufficiently large but values of length approximately 80 and 400, respectively, were optimal in this experiment. The optimal filter length corresponds to the recovered source signal to interference ratios (SIRs) are optimal. In the case of the filter lengths are not large enough or too large, the SIRs will be decreased.

Table 1. The comparison between the results achieved by using different contrast functions.

Function	SIR1(*dB*)	SIR2(*dB*)
G_1	10.2	11.6
G_2	10.5	12.3
G_3	12.6	14.1

Table 2. The comparison between the convergence speed of the innovation extraction stage.

Function	Source 1 (iterations)	Source 2 (iterations)
G_1	31	36
G_2	61	44
G_3	25	27

Table I shows the recovered SIRs in the first experiment. In this case, the criterion approximating negentropy by G_3 turned out to yield better result and indicate a good separation. This result is depicted in Fig.5. The true contributions of the speech signals is shown in Fig.6. The convergence speed of the innovation extraction stage is shown in Table II. We also used the Tugnait's method [1] in this case and it requires more than 3000 iterations to converge. We extend this experiment with 20 different sets of simulated 64th-order mixing filters (*headmix.m* in [18]). In this experiment, approximating negentropy by G_3 (kurtosis criterion) be the best optimization criterion and yield good separation performances with the mean SIR1 were 12.4-dB and SIR2 were 14.5-dB. The remain contrast functions yield lower SIRs but better robustness. In particular, these criterion are more robust to extreme than the kurtosis criterion, which involves a fourth-order moment, whose estimation is sensitive to outlier.

We also tested our above experiment with 20 different sets of simulated 256th-order mixing filters. In this experiment, the recovered source SIRs were varied from 7.2- to 11.7-dB. In the case of using G_3, the mean SIRs were 11.1 -dB for the first speech source and 11.5-dB for the second source.

The next experiment was implemented to test the method's ability in highly reverberant conditions in case of $N = 2$. To do this, we used Alex Westner's room acoustics data which have substantial reverberation for hundreds of milliseconds [19]. In this case, because the iterative rule for FIR-filter learning is complicated, the method is impossible to separate speech signals.

The last experiment was implemented to test the method's ability in case of $N = 3$. To do this, we used random sets of mixing filters of which filter orders vary from 3 to 12. However, the short filters used in this case are far from the dense impulse responses often met in realistic acoustic environments. We performed this experiment with 20 different sets. We chose the length of the inverse and re-coloration filters approximately 30 and 100, respectively. In this experiment, the recovered source SIRs were varied from 6.7- to 8.1-dB and the mean SIRs using G_3 were about 8-dB. In this case, the method with the deflation

scheme provide lower SIRs perhaps because the estimation errors in the sources that are estimated first accumulate and increase the errors in the later estimated sources.That is the reason signal deflation-based methods are sometimes unables to extract more than two sources from a multi-source mixture.

From experimental results, it is known that this proposed method (especially using G_3) can achieve a good separation performance only in the case of mixtures with short-tap FIR filters (under artificial or short reverberant conditions). Moreover, note that we assume that the sources are stationary, which implies that this method may not be the most suited for speech separation under real acoustic environments. Despite of the above limitations, we can apply this proposed method to separate speech signals in some restricted cases or to improve speech separation performances in highly reverberant conditions. In [21], authors proposed the MultistageICA combining Frequency Domain (FD)-ICA and Time Domain (TD)-ICA. In the first stage, we perform FD-ICA to separate the source signals. In the second stage, we regard the separated signals of FD-ICA as the input signals for TD-ICA and we remove the residual crosstalk components of FD-ICA by using the proposed method. Finally, we regard the output signals of TD-ICA as the resultant separated signals.

We can also use this method for telecom signals (the typical orders of the mixing filters encountered in telecommunications are more adapted to this method) or images in some restricted areas (microscopy, tomorgraphy, …).

Finally, in this paper, we assume the noise in (1) is negligible so a main disadvantage of this method is the lack of any analysis of the effects of noise. With the existense of noise, the model in (1) becomes the underdetemined case and the proposed method doesn't work well. The ICA based methods are very strongly effected in noise but an investigation of such a model is however beyond the scope of this paper.

IV. CONCLUSIONS

In this paper, we have proposed the approach extended from [1-4], which combines inverse filter criteria with negentropy maximization to separate convolutive mixtures of speech sources in the time domain. Sufficient conditions for separating speech sources has established. The limitations of the proposed approach in separation of speech sources have also demonstrated. One of the strong point of this approach is that the model order needs not be known as long as extraction and re-coloration filters are "long enough". The limitation of our research is the lack of any comparision of the proposed method with others since the other time domain ICA algorithms are not available either in internet or under request.

REFERENCES

[1] J.K.Tugait, "Identification and deconvolution of multichannel linear non-Gaussian processes using higher order statistics and inverse filter criteria", IEEE Transactions on Signal Processing, Vol.45, No.3, March 1997.

[2] C.Simon et al, "Blind source separation of convolutive mixtures by maximization of fourth order cumulants: the non-iid case, "Proceedings of The Thirty-Second Asilomar Conference on Signals, Systems & Computers, November 1998, Vol.2 , pp1584-1588.

[3] F.Abrard et al., "Blind source separation in convolutive mixtures:a hybrid approach for colored sources", IWANN 2001, LNCS2085, pp 802-809, 2001.

[4] Johan Thomas et al "Time Domain Fast Fixed Point Algorithms for Convolutive ICA", IEEE Signal Processing Letters, Vol.13, No. 4, April 2006.

[5] L.R.Rabiner and R.W.Schafer, "Digital Processing of Speech Signals", Prentice-Hall, Upper Saddle River, NJ, USA, 1983

[6] Monson H.Hayes, "Statistical Digital Signal Processing and Modeling", John Wiley & Sons, Ltd, 1996.

[7] K.Kokkinakis and A.K.Nandi, "Multichannel blind deconvolution for source separation in convolutive mixtures of speech", IEEE Transactions on Audio, Speech and Language processing, Vol.14, No.1, January 2006.

[8] A.Cichocki et al, "A blind extraction of temporally correlated but statistically dependent acoustic signals",

Neural Network for Signal Processing, X, 2000, Proceedings of the 2000 IEEE Signal Processing Society Workshop, Vol.1, pp.455-464.

[9] T.Yoshioka et al, "Dereverberation by using time-variant nature of speech production system", EURASIP Journal on Advances in Signal Processing, vol.2007.

[10] A.Kizilaya et al "Estimation of the ARMA model parameters based on the equivalent MA approach", The second IEE-EURASIP Int.Symp.on Communi-cations, Control and Signal processing, ISCCSP'06 Marrakech, Marocco, 2006.

[11] A.Hyvarien, "Fast and robust fixed-point algorithms for independent component analysis", IEEE Transaction on Neural Networks, 10(3):626-634, 1999.

[12] Aapo Hyvarinen et al, " Independent component analysis: Algorithms and Applications", Neural Networks, 13(4-5):411-430, 2000.

[13] F.Abrard et al, "Blind partial separation of underdetermined convolutive mixtures of complex sources based on differential normalized kurtosis", Elsevier, Neurocomputing 71(2008), pp 2071-2086.

[14] N. Delfosse and P. Loubaton, "Adaptive blind separation of convolutive mixtures", ICASSP'96: Proceedings of the Acoustics, Speech, and Signal Processing, 1996 IEEE International Conference, Vol.5, pp.2940-2943.

[15] N.Mitianoudis and M.E.Davies, "Audio source separation of convolutive mixtures", IEEE Transactions on Speech Audio Process, Vol.11, No.5, pp489-497, Sep.2003.

[16] J.Thomas, Y.Deville, S.Hoseini "Differential fast fixed-point algorithms for underdetermined instantaneous and covolutive partial blind source separation", IEEE Transactions on Signal Processing, Vol.55, No.7, July 2007.

[17] Lang Tong, "Identification multichannel MA parameters using higher order statistics", Elsevier, Signal Processing 53 (1996), pp 195-209.

[18] http://sound.media.mit.edu/ica-bench/

[19] http://www.media.mit.edu/~westner

[20] J.Thomas, Y.Deville, S.Hoseini, "Fixed-point algorithms for convolutive blind source separation based on non-gaussianity maximization", Proceedings of the 7th International Workshop ECMS'05, Toulouse, France, May 2005.

[21] T. Nishikawa, H. Saruwatari, and K. Shikano, "Blind Source Separation Based on Multi-Stage ICA Combining Frequency-Domain ICA and Time-Domain ICA", Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2002), pp.2938--2941, May 2002.

AUTHORS'S BIOGRAPHIES

Nguyen Quoc Trung was born in 1949 in Nam Dinh, Vietnam. He received the Ph.D in 1982 and was promoted to Associate Professor in 2004. He is currently a Lecturer in the Faculty of Electronics and Telecommunications, Hanoi University of Science and Technology. His professional research interests are digital signal processing, filter theory.

Tran Hoai Linh was born in 1974 in Hanoi, Vietnam. He received the M.Sc. in Applied Informatics, Ph.D and Dr.Sc. in Electrical Engineering from the Warsaw University of Technology in 1997, 2000 and 2005, respectively. He was promoted to Associate Professor in 2007. He is currently a Researcher and Lecturer in the Department of Instrumentations and Industrial Informatics, Faculty of Electrical Engineering, Hanoi University of Science and Technology. His professional research interests are artificial methods and applications in classification and estimation problems.

Vuong Hoang Nam was born in 1980 in Hanoi, Vietnam. He received the M.Sc in 2005 in Hanoi University of Technology. He is currently a Lecturer in the Faculty of Electronics and Telecommunications, Hanoi University of Science and Technology. His professional research interests are digital signal processing, multimedia applications.