ESTIMATION OF SPEECH EMBEDDED IN A REVERBERANT ENVIRONMENT WITH MULTIPLE SOURCES OF NOISE

Allan Kardec Barros[1,4]*, Fumitada Itakura*[2]*, Tomasz Rutkowski*[3]*, Ali Mansour*[1] *Noboru Ohnishi*[1,2]
1 : BMC, RIKEN, Japan. 3 : BSI, RIKEN, Japan.
2 : CIAIR, Nagoya University, Japan. 4 : UFMA, Brazil.
E-mail: akbarros@ieee.org.

ABSTRACT

In this work we develop a system for enhancement of the speech signal with highest energy from a linear convolutive mixture of n statistically independent sound sources recorded by m microphones, where $m < n$. In this system we use the concept of independent component analysis (ICA) together with auditory filter banks, pitch tracking, adaptive band pass filters and masking. Computer simulations and real world experiments confirm the validity of the proposed algorithm.

1. INTRODUCTION

The *cocktail party problem* is well known in auditory scene analysis: in a room there are many sources of sound mixed and reverberated: voice, music, noise, etc. The task is to segregate one or more of those sound signals and improve their intelligibility.

Independent component analysis (ICA) appears as an important technique to help solving this problem. This is because ICA algorithms find a linear combination of the mixed signals which recovers the original (or source) signals, possibly re-scaled and randomly arranged in the outputs.

However, there are at least two difficulties related with the *cocktail party* problem: firstly, due to reverbation effect, we actually observe a convolutive mixture; secondly, in practice we have smaller number of microphones than unknown acoustic source signals, thus standard ICA cannot be directly applied. It is important to notice that humans can deal with this problem by using only two ears.

Similarly to humans, our aim here is not to recover simultaneously all the original acoustic signals. Our task is rather to turn a specific speech signal more intelligible than the available microphone signals. As does our auditory system, we try to enhance the signal nearest to the microphones, i.e., the signal with highest energy. We realize this by mimicking some properties of human auditory system. This is carried out by (a) mimicking the inner ear, through the use a bank of self-adaptive band-pass wavelet filters (b) the tracking of the speech fundamental frequency $f0$ and; (c) by masking some parts of the speech with lower energy.

There are two contributions that we find important in this manuscript: one is that we propose an algorithm which at the same time extracts pitch and decides whether the current part of speech is voiced or not; and the other is that we propose an ICA algorithm which, contrary to other works, e.g., [18, 3], outputs only one signal, through the use of $f0$ information. Thus, the so-called *permutation problem* [9] in ICA is solved.

2. THE METHOD

Consider n source signals at time t, $\mathbf{s} = [s_1(t), s_2(t), \ldots, s_n(t)]^T$ arriving at m receivers $\tilde{\mathbf{x}}(t) = [\tilde{x}_1(k), \tilde{x}_2(t), \ldots, \tilde{x}_m(t)]^T$. In the linear model of cocktail party, each receiver gets a combination of the source signals, so that we have,

$$\tilde{\mathbf{x}}(t) = \int_{-\infty}^{t} \mathbf{H}(\tau)\mathbf{s}(t+\tau)d\tau \qquad (1)$$

where $\mathbf{H}$ is a linear filter operator, which models the reverberation and mixing. It is important to notice that in an actual environment, $\mathbf{H}$ is a non-minimum phase low-pass filter [11], which turns the task of recovering the original signals very difficult.

In this work, we employ many features of human auditory system, in the way shown in Fig. 1. Firstly we track the pitch through an algorithm called speech instantaneous frequency (SIF). Secondly, we adaptively filter the corrupted signal $\tilde{\mathbf{x}}$, in different sub-bands using $[f0, 2f0, ...]$ as the central frequencies. Then, we take each sub-band output and enter them in an ICA algorithm whose task is to find the signal which is mostly related to $f0$ and its multiples.

Our final objective is to develop an algorithm whose output signal $y(t)$ is a modified version of a given source signal $s_i(t)$, i.e., the signal of interest will be given by $y(t) = g(s_i(t))$, where $g(\cdot)$ can at the same time be a filter and a non-linear transformation operator.

Also in our algorithm we included the temporal masking characteristic of the auditory system. This is managed by a switch which is one for the voiced part, and decays gradually to zero in the silent and unvoiced part. This switching

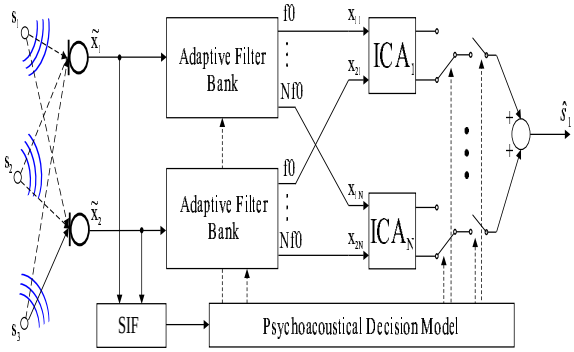

Fig. 1. Block diagram of the algorithm which mimics the auditory system. First ly it tracks the fundamental frequency (f0) using SIF. Then, it process the mixed signals using a bank of band-pass filters (such as the inner ear). After, it process each mixed/reverberated signal by an ICA algorithm. Finally, it carries out masking by turning switches *on* or *off*.

is managed by a variable estimated by SIF, which we call *speech instantaneous amplitude*, as we shall see in the next section.

2.1. Extraction of the Fundamental Frequency

The extraction involves firstly the estimation of the spectrogram which, for a signal $\tilde{x}(t)$, is defined as

$$P(t,f) = \frac{1}{2\pi}\left|\int e^{-j2\pi f\tau}\tilde{x}(\tau)h(\tau - t)d\tau\right|^2. \qquad (2)$$

where $h(\tau - t)$ is a window function.

After this, we look for the frequency value corresponding to the maximum of $P(t,f)$ at each time instant in a given frequency range. We call this quantity the *driver* $\delta(t)$, and it is given by

$$\delta(t) = \arg\max_f \left[P(t,f)\right]_{\delta(t^-)-\alpha}^{\delta(t^-)+\alpha}. \qquad (3)$$

We define α as a frequency value that limits the searching range, and $\delta(t^-)$ as the driver value at the previous time t^-. Generally speaking, we wanted to say with (3) that the algorithm searches for the maximum of $P(t,f)$ at each time instant t, along the frequency axis f which is bounded to interval $[\delta(t^-) - \alpha, \delta(t^-) + \alpha]$.

After this, we calculate the instantaneous frequency by using a band-pass filter around a central frequency given at each time instant by the *driver*. In particular, we use wavelets to construct the filter. The basic wavelet is a slight modification of the Gabor function, which is localized in both time and frequency domains. The modification is carried out in order to shift the spectral response of the filter to the central frequency. Thus, the basic wavelet is [5]

$$\Psi(t) = \frac{1}{2\pi}\frac{d}{dt}\left[e^{-\pi\left\{\frac{\overline{\delta(t)}t}{10}\right\}^2}\cos\left(2\pi t\int_\Omega \delta(\tau)d\tau\right)\right],$$

$$\overline{\delta(t)} = \frac{1}{\Omega}\sum_\Omega \delta(t). \qquad (4)$$

where Ω is a short time interval[1]. The signal filtered in this interval is given by

$$r_\Omega(t) = \int_\Omega z_\Omega(\tau)\psi(t-\tau)d\tau. \qquad (5)$$

The speech instantaneous frequency is then calculated substituting (5) into the following equation,

$$\omega_\Omega(t) = \frac{d\phi_\Omega(t)}{dt}, \qquad \phi_\Omega(t) = \arctan\left(\frac{-H[r_\Omega(t)]}{r_\Omega(t)}\right) \qquad (6)$$

where $H[s(t)]$ is the Hilbert transform of the signal $r_\Omega(t)$.

Another variable that we extract from the signal $r_\Omega(t)$ is the speech instantaneous amplitude, given by $\alpha_\Omega(t) = |H[r_\Omega(t)]|$. This term is responsible for the switching at the last step of the algorithm.

2.2. The Bank of Adaptive Band-pass Filters

We use here the concept of harmonicity of the voiced sounds that was exploited in some models of *computational auditory scene analysis* (CASA) which group together spectrotemporal regions that are modulated by the same period [20].

The idea is to use a bank of band pass filters centered at the fundamental frequency $f0$ and its harmonics, as proposed in [5]. We use the same wavelet as given in (4), but now we substitute $\delta(t)$ by the estimated $f0(t)$.

From this, we obtain intermediary signals $r_{i,k}(t)$ which are $\tilde{x}_i(t)$ filtered around frequency $kf0(t)$, $(k = 1, 2, \ldots, N)$, given by

$$r_{i,k}(t) = \int_{-\infty}^{\infty}\Psi(t,k)\tilde{x}_i(t)d\tau$$
$$\text{for } k = 1,\ldots,N,\ i = 1,2,\ldots,m. \qquad (7)$$

where N is the number of harmonics (and therefore of subbands).

Then, we find the instantaneous amplitude of each $r_{i,k}(t)$ by the following operation,

$$\hat{a}_{i,k}(t) = |H[r_{i,k}(t)]| \qquad (8)$$

where $H[r_{i,k}(t)]$ is the Hilbert transform of the signal $r_{i,k}(t)$.

At this point, however, we have no phase information about the signal we want to estimate. Thus, we generate from $\hat{a}_{i,k}(t)$ and $f0(t)$ a set of orthogonal signals,

$$z_{q,i,k} = \hat{a}_{i,k}(t)e^{qj2\pi tkf0(t)}, \qquad (9)$$
$$q = -1, 1 \text{ and } k = 1,\ldots,N.$$

[1] We used zero padding to handle border distortions caused by the use of wavelets.

In order to obtain the phase information of the signal, we use the Wiener theory. In this case the $i-th$ output of the k-th sub-band will be,

$$x_{i,k} = \mathbf{c}_{i,k}^T \mathbf{z}_{i,k}(t),\ i = 1, 2, \ldots, m. \quad (10)$$

where $\mathbf{z}_{i,k}(t) = [z_{1,i,k},\ z_{-1,i,k}]^T$. In Wiener theory, given the signal $\hat{x}_{ik}(t)$, the weight vector $\mathbf{c}_{i,k}$ which gives the minimum mean squared error between the estimated signal $p_{i,k}$ and $\hat{x}_{ik}(t)$ is given by[12],

$$\begin{aligned} \mathbf{c}_{i,k} &= \mathbf{R}^{-1}\mathbf{P} \\ &= E[\mathbf{z}_{i,k}(t)\mathbf{z}_{i,k}(t)^T]^{-1} E[\hat{x}_{ik}(t)\mathbf{z}_{i,k}(t)]. \end{aligned} \quad (11)$$

Since the elements of $\mathbf{z}_{i,k}(t)$ are mutually orthogonal, matrix $\mathbf{R}$ is diagonal. Thus, it is not difficult to remove the inversion in (11), by normalizing the elements of $\mathbf{z}_{i,k}(t)$ to have unity variance. In this case, $\mathbf{R} = \mathbf{I}$, thus, $\mathbf{c}_{i,k} = E[x_i(t)\mathbf{z}_{i,k}(t)]$.

3. INDEPENDENT COMPONENT ANALYSIS

In this section we study the third step of the algorithm, now that we have available the sub-band signals, given by (10), obtained from the bank of band-pass filters. In other words, we have split wide-band into narrow-band signals. An important property of a narrow band signal is that they have less effects of convolution. In fact, the convolutive mixture turns approximately into an instantaneous mixture, as the bandwidth diminishes. Another important poit is that we effectively reduce the probability of finding more than two strong signals at the same sub-band.

An important contribution of this manuscript is that we no longer extract two (or more) components such as in previous works, e.g., [18, 3]. Notice also that the inputs of the ICA blocks are the outputs of the bank of band pass filters (see Fig. 1). In this case, the ICA inputs for the k-th sub-band are signals composing vector $\mathbf{x}_k$. In order to simplify notation, let us consider output of the k-th sub-band at the j-th step, and its corresponding error respectively defined as:

$$y_{j,k}(t) = \mathbf{w}_{j,k}^T \mathbf{x}_k(t),\ \varepsilon_{j,k}(t) = \mathbf{w}_{j,k}^T \mathbf{x}_k(t) - b y_{j,k}(t-p), \quad (12)$$

where $\mathbf{w}_{j,k} = [w_{11}, w_{12}, \ldots w_{im}]^T$, $\mathbf{x}_k = [x_{11}, x_{12}, \ldots x_{1m}]^T$, p is a given time delay, and b is a scalar weight. For simplicity, we will drop the time index t and make $\mathbf{y}_{j,k,p} = \mathbf{y}_{j,k}(t-p)$.

The cost function $\xi_{j,k} = E[\varepsilon_{j,k}^2]$can be evaluated as follows:

$$\xi_{j,k} = \mathbf{w}_{j,k}^T E[\mathbf{x}_k \mathbf{x}_k^T]\mathbf{w}_{j,k} - 2bE[y_{i,p}\mathbf{w}_{j,k}^T \mathbf{x}_k] + b^2 E[\mathbf{y}_{j,k,p}^2]. \quad (13)$$

In order to estimate the weight vector $\mathbf{w}_{j,k}$ we evaluate the gradient of the cost function as follows:

$$\frac{\partial \xi_{j,k}}{\partial \mathbf{w}_{j,k}} = 2E[\mathbf{x}_k \mathbf{x}_k^T]\mathbf{w}_{j,k} - 2bE[y_{i,p}\mathbf{x}_k] + 2b^2 E[\mathbf{x}_{i,p}\mathbf{x}_{i,p}] \quad (14)$$

Solving $\frac{\partial \xi_{j,k}}{\partial \mathbf{w}_{j,k}} = \mathbf{0}$ we obtain a new iterative algorithm given by,

$$\mathbf{w} = E[\mathbf{x}_k \mathbf{x}_k^T]^{-1} E[y_{i,p}\mathbf{x}_k] \frac{b}{1+2b^2}, \quad (15)$$

In order to avoid the trivial solution $\mathbf{w}_{j,k} = \mathbf{0}$, we perform normalization of the vector to unit length at each iteration step as $\mathbf{w}_{j,k,*} = \mathbf{w}_{j,k}//||\mathbf{w}_{j,k}||$. With this, the term $b/(1+2b^2)$ can be disregarded. Moreover, we can assume without loss of generality that the sensor data are prewhitened, thus $E[\mathbf{x}_k \mathbf{x}_k^T] = \mathbf{I}$. With this, (15) leads to a very simple learning rule,

$$\mathbf{w} = E[\mathbf{x}_k y_{i,p}]. \quad (16)$$

In a previous work, Barros and Cichocki [4], have studied the properties of the above algorithm. One of them is that, if the signals are mutually independent and if for one of the source signals, say s_j, the autocorrelation property $\mathbf{s}_j(t)\mathbf{s}_j(t-p) \neq 0$ holds, then the algorithm output will be $\mathbf{s}_j$ up to a scaling factor [2]. Thus, since in a previous step we have estimated the voice pitch $f0$, we can easily use as the necessary delay $p = 1/f0$, for the first frequency band $i = 1$ and for the other bands we just make $p = 1/(k.f0)$.

Although we solved the permutation problem, the *scaling* effect well known in ICA [9], persists. For this, we use an exponential spectral decay, i.e., $\hat{y}_{j,k,p} = e^{-\alpha k} y_{j,k,p}$, where $0 < \alpha < 1$.

4. RESULTS

Firstly, we carried out simulations where we mixed and convoluted three independent speech signals into two mixtures, as modeled by (1), where $n = 3$ and $m = 2$. The *desired* signal was a male voice and the interferences were a male singing and the sound of a laugh. The coefficients of the convolution filter were 100, whereas we turn some of them zero to roughly mimic a real room impulse response. The task for the algorithm, as we have stated before, was to find the signal with the highest energy.

Similarly, we have carried out real world experiments in a standard laboratory room ($5m \times 7m$), where we placed two microphones in the middle, away 1.5 m from each other. In the environment, there were computers, tables, chairs, etc. One male speaker stood up in front of them, away 1.5 meters. Behind him, making a triangle with height of 1.5 m, was a female speaker and a speaker phone playing music. The data was sampled at 16 KHz and recorded on a personal computer.

As we have shown before, the output signal is no longer a linearly mixed non-delayed version of the original source

[2] This is demonstrated in[4].

signal, therefore we cannot measure easily how much background noise was removed from the mixed signal, or how distorted the source came out, by means of a simple technique such as mean squared error. Thus, we have opted for the subjective measurements by the MOS scale [7], which is a five-point rating scale, covering the options Excellent, Good, Fair, Poor and Bad. Ten subjects were asked to rate separately: a) The background noise and; b) The distortion introduced by the algorithm. Each sound was played twice in random order using the MATLAB command *sound*. The results are shown in Table 1.

BACKGR. NOISE SENS.	*Simulation*	*Real World*
Conv./Mixed	1.75	2.02
Output	3.89	3.32

Table 1: The mean MOS score accounting for the background noise sensitivity at each stage of the process.

As expected, the system worked more efficiently in the case of simulation than in the real world experiment. This is explained by the fact that we do not know the impulse response of a real room, including its non-minimum phase effect. On the other hand, while in the simulation the mixing was generally evaluated as *poor* by the listeners and the output *good*, in the real room case there was only one step improvement from *poor* to *regular*. This may be explained by the fact that some aliasing should be occurring when the sub-bands are added.

One can see that pitch tracking, bank of filters and the assumption of statistical independence of the sources are encouraging. Compared to previous works[3], where the main focus was that of when the number of sources and sensors are the same, in this paper we went one step ahead and made the number of sources greater than the number of mixtures, and turned the problem difficult to be solved by the algorithms proposed until now.

5. REFERENCES

[1] S. Amari and A. Cichocki: "Adaptive blind signal processing - neural network approaches," *Proceedings IEEE* (invited paper), Vol.86, No.10, Oct. 1998, pp. 2026-2048.

[2] S. Amari, "ICA of temporally correlated signals - learning algorithm," *Proc. ICA '99*, Aussois, France, pp. 13-18, Jan. 1999.

[3] A. K. Barros., H. Kawahara, A. Cichocki, S. Kajita, T. Rutkowski, M. Kawamoto, N. Ohnishi. Enhancement of a Speech Signal Embedded in Noisy Environment Using Two Microphones. *Proc. ICA'2000*, Helsinki, Finland. v.1. pp.423-428.

[4] A. K. Barros and A. Cichocki, ". Extraction of Specific Signals with Temporal Structure" accepted for publication by Neural Computation.

[3] Further references can be found in the ICA'99 and ICA'2000 proceedings.

[5] A. K. Barros and N. Ohnishi, "Amplitude estimation of quasi-periodic physiological signals by wavelets", accepted for publication by IEICE.

[6] A. Belouchrani, K. Meraim, J.-F. Cardoso and E. Moulines, "A blind source separation technique based on second order statistics". *IEEE Trans. on Signal Processing*, 45, pp. 434-444, 1997.

[7] CCITT, Recommendations of the P Series, "Method for the evaluation of service from the standpoint of speech transmission quality". CCITT Red Book Volume V - VIIIth Plenary Assembly, 1984.

[8] A. de Cheveign, "The auditory system as a separation machine", *Proc. International Symposium on Hearing*, in preparation, 2000.

[9] P. Comon, (1994) "Independent component analysis, a new concept?" *Signal Processing*, 24, pp. 287 - 314.

[10] N. Delfosse and P. Loubaton, "Adaptive blind separation of independent sources: a deflation approach", *Signal Processing*, vol. 45, pp. 59 - 83, 1995.

[11] B. Gold and N. Morgan. *Speech and audio signal processing.* John Wiley and Sons, 2000.

[12] S. Haykin, *Adaptive filter theory.* Englewood Cliffs, NJ: Prentice-Hall, 1991.

[13] A. Hyvarinen and E. Oja, "A fast fixed-point algorithm for independent component analysis". *Neural Computation* (9), 1483 - 1492, 1997.

[14] C. Jutten and J. Hérault "Independent component analysis versus PCA," *Proc. EUSIPCO*, pp. 643 - 646, 1988.

[15] H. Kawahara, I. Masuda-Katsuse and A. de Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds", *Speech Communication*, 27, pp.187-207 1999.

[16] T-W Lee, *Independent component analysis.* Kluwer Academic Publishers, 1998.

[17] L. Molgedey, H.g. Schuster, "Separation of a mixture of independent signals using time-delayed correlations, *Phys. Rev. Lett.*, vol. 72(23), pp. 3634-3637, 1994.

[18] S. Ikeda and N. Murata, "A method of ICA in time frequency domain," *Proc. ICA '99*, Aussois, France, pp. 365-370, Jan. 1999.

[19] A. Papoulis. *Probability, random variables, and stochastic processes.* McGraw-Hill, 1991.

[20] Weintraub, M. "A theory and computational model of auditory monaural sound separation", Doctoral dissertation, Stanford University.1985

[21] K-C Yen and Y. Zhao, (1990). "Adaptive co-channel speech separation and recognition", *IEEE Trans. on Speech and Audioprocessing*, vol. 7, No. 2, pp. 138-151, 1999.

[22] Zissmann, M. A., and Weinstein, C. J. (1990). "Automatic talker activity labeling for co-channel talker interference suppression.", Proc. IEEE-ICASSP, 813-816.

[23] Virag, N., "Single channel speech enhancement based on masking properties of the human auditory system". *IEEE Trans. on Signal Processing*, Vol. 7, No. 2, pp. 126-137, 1999.