REAL TIME SPEECH SEPARATION BY LATERAL INHIBITION AND MASKING

Allan Kardec Barros[1], Edil James[1], Yoshinori Takeuchi[2]

[1] Department of Electrica Engineering
Universidade Federal do Maranhão
São Luís-MA, Brazil
[2]Graduate School of Information Science
Nagoya University, Nagoya 464-8603, Japan

Abstract: In this paper, we propose a simple algorithm to separate a speech signals with the highest energy from a mixture of sound sources . We use two microphones and assume that one speaker is close to one microphone, and the other speaker is close to another microphone. In the system we use the concept of auditory filter banks together with lateral inhibition, intensity interaural difference and masking. This algorithm is so simple that we can easily implement as a real-time speech separation system. Computer simulations and real world experiments confirm the validity of the proposed algorithm.

Keywords: cocktail party, auditory filter banks, interaural intensity difference, lateral inhibition and masking

1. INTRODUCTION

Among the problems in auditory scene analysis, perhaps the widest known is the cocktail party, which is generally related to the problem of selective attention, how humans can select the voice of a particular speaker in a noisy environment wher there are many sources of sound mixed and reverberated: voice, music, air-conditioning noise, etc. The task is to segregate one or more of those sound signals, or enhance their intelligibility. A number of solutions were proposed to solve this matter. Some involved the use of the harmonicity characteristic of human speech, through its fundamental frequency (Parsons, 1976; Aoki, et al., 2001), by subtractive -type algorithms (Virag, 1999), or through independent component analysis (ICA) (Barros, et al., 2002).

However, in situations such as a conversation carried out between robot and humans, the sound separation must work in real-time and we cannot use time-consuming algorithms in this application. Therefore, a computationally fas algorithm is required.

There are many problems involved in the cocktail party. Firstly, the mixtures arriving at the microphones are reverberated versions of the original source. Secondly, the room impulse response changes according to parameters such as the distribution of furniture in the room, wall material or temperature. Another problem is that the target is not usually static, and some of those models proposed in the literature may fail. It is interesting that human deal with this matter using only two ears, this occurs becau se the sounds are filtered by thousands of band-pass filter in the cochlea.

In this paper, we propose a very simple on-line algorithm to separate two speech signals. We use two microphones, and assume that one speaker is close to one microphone, and the other speaker is close to the other one. Then, separation is carried out by separating the signals in different frequency bands and comparing the power of the corresponding frequency component. As does our auditory system, we try to enhancement the signal nearest to the microphones, i.e., the signal with highest energy. We realize this by mimicking some properties of the human auditory system.

2. HUMAN AUDITORY PERCEPTION

Our algorithm is highly motivated by human auditory perception. Thus, in this section, e describe the binaural hearing briefly.

The cochlear duct together with the basilar membrane of the ear work as a frequency analyzer. Thus, the earlier stages of the auditory system may be understood as a bank of band-pass filters with frequency overlapp ing their neighbours.

Some interesting phenomena occu at the *auditory cortex* as the *lateral inhibition* and *masking*. They can be understood as “fine adjustment” because they help in the selectivity of desired signal. Lateral inhibition is a process in which the signal with higher energy inhibits the other at some stage of the auditor processing. Masking appears as an important tool of the human hearing system. It is the process through which the threshold of audibility of a sound is shifted in the presence of another sound, which means that a sound masked by anothe is difficult or impossible to be heard (Moore Brian, 1997).

In a real environment, the times of arrival of sounds at left and right ears are different. Therefore, it was coined by the research community the term *interaural time difference* (ITD) to designate that difference, which is very useful to localize the sound source. Interestingly, higher frequency components ar attenuated by head that creates a barrier causing an acoustic shadow, so that there is as well the so-called interaural intensity difference (IID) between left and right ears. As the calculation of IID is easier than that of ITD, we propose to use the IID in our algorithm.

We use these three concepts in this work.

3. SEPARATION ALGORITHM AND ITS IMPLEMENTATION

3.1 Separation Algorithm

We assume that source signals are sparse in frequency domain, i.e., a certain frequency band of mixed signals has only one source signal. The idea is that, if two speakers are different, their fundamental frequencies are different as well.

Let us consider the two speakers and two microphone case. If speaker A is close to microphone A, the other speaker B must be close to the other microphone B. In this case, the speech signal from the speaker A obtained at the microphone A is larger than that obtained at microphone B and the speech signal from the speaker B obtained at the microphone A is smaller than that obtained at the microphone B. Adding to this, the concept of lateral inhibition and auditory masking in a given frequency band, we can only actually hear on of them, although the other sound may be present. Therefore, we can separate the source signals based on the intensity of the observed signals.

Our separation algorithm is very simple as show in figure 1. Lets x1(t), x2(t) be mixed signal observed at two microphones. At first, each input signal is filtered, at the bank of band-pass filter, in different sub bands, using [f0, 2f0, ..., nf0] as the central frequencies.

$$\begin{aligned} f_{1i}(t) &= BPF_i * x_1(t) \\ f_{2i}(t) &= BPF_i * x_2(t) \end{aligned} \qquad (1)$$

where BPFi is a band-pass filter and * denotes convolution operator

Then, we take each sub band output and enter them in a *lateral inhibition*, which compare the power of each filtered signal within same filter bank, selecting the bands of larger level of energy and inhibiting their closer neighbours.

$$g1i(t =) \begin{cases} f1(i\text{-}1)=0 & \\ f1j(t) & \text{if } E(f1j(t)) > E(f1i(t)) \\ f1(i+1)=0 & \end{cases}$$

$$(2)$$

$$g2i(t =) \begin{cases} f2(i\text{-}1)=0 & \\ f2j(t) & \text{if } E(f2j(t)) > E(f2i(t)) \\ f2(i+1)=0 & \end{cases}$$

where *E(.)* is an envelope estimato , *gi(t)* is a train pulse formed by the bands of larger energy of the input signal. Also in our algorithm we included the *temporal masking* characteristic of the auditory system. This is managed by comparing the magnitude of the bands with the same central frequency, by a switch to the one which is for the signal of larger energy of the chosen microphone and zero for the others. Finally, we add the output signals y(t) to create the separated signal:

$$y(t) = \sum_{i=1}^{M} wi(t) \qquad (3)$$

where M is the number of filters and *wi(t)* is a sample of signal recovered that correspond the *fi(t).*

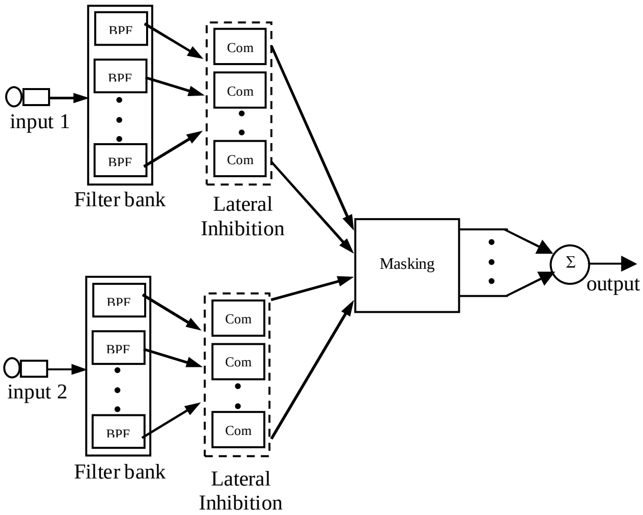


Fig. 1. Block diagram of the algorithm which mimic the auditory system.

Since our algorithm is very simple, calculation time is very short and can be easily implemented in real-time, thus yielding a great advantage over others.

4. EXPERIMENTAL RESULT

Firstly, we carried out simulations where we mixed and convoluted two signals into two mixtures. The desired signal was a male voice and the interference as a female voice.

The task was to find the signal with the highest energy. These simulations aimed to mimic the case when one speaker is close to the listener, but there is some background interference.

Figure 2 and 3 show experimental results obtained to recover the original signal, using only lateral inhibition module and it together with masking respectively. In both simulations, tests accomplished used the simulated mixed fo computer and mixed obtained in real world.

The figure, s1 and s2 are two source signals, x and x2 are the signals obtained by two microphones, and y is the output signal. We used two Portuguese male and female utterances as source signals. The reverberation time in the reverberant room was 0.5s.

We have measured the MOS scale [CCITT, Recommendations, 1984] to give the subjective evaluation. We used the typical MOS which is a 5-point rating scale, covering the options excellent, Good, Fair, Poor and Bad. Ten subjects are asked separately: 1) How much can you hear the interference signal? and; 2) Resulting quality. Each sound was played twice in random order. The results are show in Table1.

We also accomplished a simulation, where we used only the lateral inhibition to separate the desired signal of the mixed. The principle is the same showed in figure1 without the masking module. The result presented a noisy that the listeners compared as the sound produced by a cricket.

Similarly, we have carried out real world experiments. The sampling rate of each input signal was 8 kHz in both experiments. We use 78 band-pass filters such that the difference among the central frequencies was 25Hz.

5. DISCUSSIONS

Since t e output signal is created by the combination of the band-pass filtered signal, some part of the desired signal may be dropped and some part of the interference signal may be added. We can find that our algorithm enhances the quality of the signal. Especially, in the reverberant room there ere two steps decrease of the quality. In a room impulse response changes according to parameters such as the distribution of furniture in the room, wall material or temperature. This may cause the quality decrease.

	Lateral Inhibition		Lateral Inhibition and Masking	
	Input	Output	Input	Output
Interference	2,0	4,0	2,0	3,0
Quality	4,0	2,8	4,0	4,0

(a) computational simulations

	Lateral Inhibition		Lateral Inhibition and Masking	
	Input	Output	Input	Output
Interference	2,5	3,5	2,5	3,0
Quality	3,5	1,5	3,5	2,0

(b) reverberant room

Table 1- The MOS score of the input and output signal.

From the figure 2, we can see the performance of the separation works well in the simulated mixture fo computer. Our algorithm does not involve the scaling and permutation problem as ICA. Thus, we can obtain the power of the each source signal.

As expected, the system worked more efficiently in the computational simulations than in the reverberant room. This is explained the fact that the reverberant waves does not meet the signal intensity assumption. On the other hand, while in the computational simulations the interference sensitivity of the input signal was generally evaluated as poor by the listeners and the output good, in the reverberant room there was only one step improvement from poor-fair to fair-good. This may be explained by the fact that some part of reverberant waves of the interference signal still remain in the output signal.

6. CONCLUSION

We proposed a simple separation algorithm that can work in real-time. Our algorithm is motived by human auditory perception, especially lateral inhibition and masking of binaural hearing. Frequency analysis of the mixed signal is carried out in cochlear duct and IID is used to separate the mixed signal.

We have implemented our algorithm in PC. Our algorithm is very suitable for online implementation an d also hardwar implementation.

We have conducted experiments in both computational simulations and real environment and showed that ou algorithm can separate mixed sources.

Further work should be carried out to improve the performance in the reverberant room. It can be realized by using a precedence effect.

7. REFERENCES

A. K. Barros, T. Rutkowski, F. Itakura, N. Ohnishi, “Estimation of Speech Embedded in Reverberant and Noisy Environment by Independent Component Analysis and Wavelets”. IEEE Trans. On Neural Networks, Vol. 13, No. 4, pp. 888 -893, 2002.

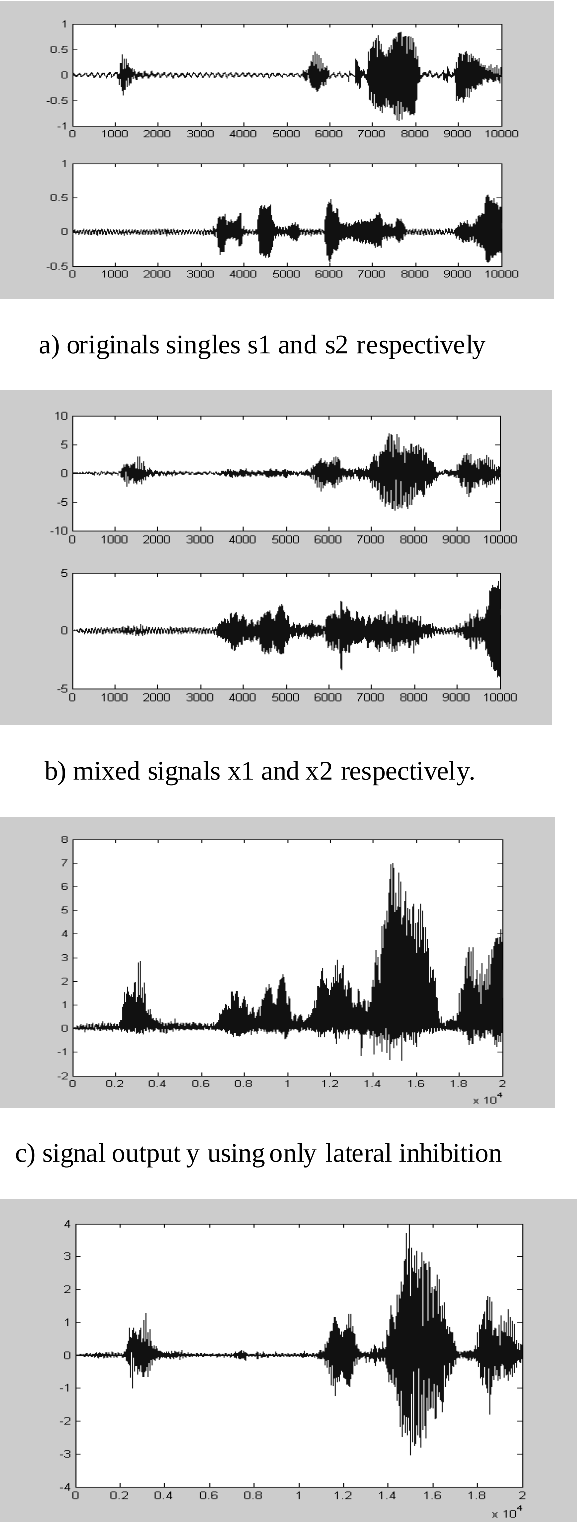
CCITT, Recommendations of the P. Series, “Method for the evaluation of service from the standpoint of speech transmission quality,” CCITT Red Book Volume V – VIIIth Plenary Assembly, 1984.

M. Aoki, M. Okamoto, S. Aoki, H. Matsui, T. Sakurai and Y. Kaneda, “Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones,” Acout. Sci. And Tech, 22, 2, pp. 149 -157, 2001.

Moore Brian C.J. “An introduction to the psychology of hearing”. Academic press 4th edition, 1997.

N. Virag, “Single channel speech enhancement based on masking properties of the human auditory system”. IEEE Trans. On Signal Processing, Vol7, No. 2. Pp. 126 -137, 1999.

T.W. Parsons, “Separations of speech from interfering speech by means of harmonic selection,” Journal of the Acoustical Society of America, 60, pp. 911-918, 197 .



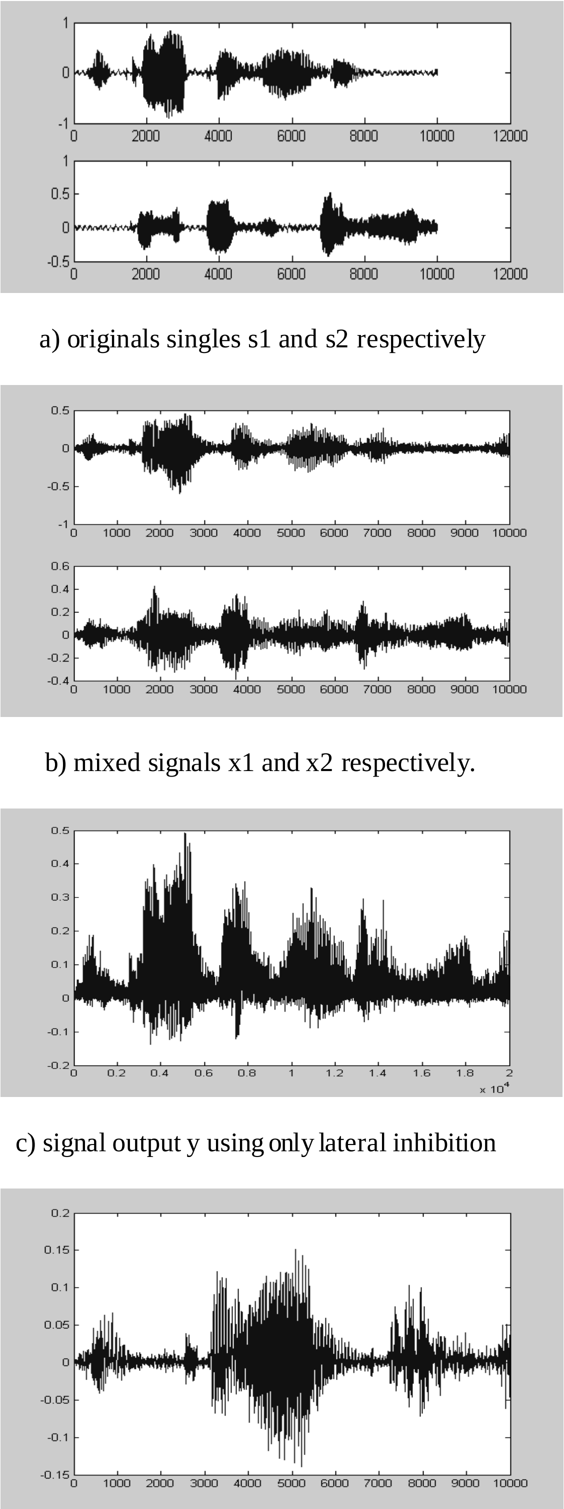
a) originals singles s1 and s2 respectively

b) mixed signals x1 and x2 respectively.

c) signal output y using only lateral inhibition

d) signal output y using lateral inhibition and masking

Fig. 2 – Signals obtained with computational simulations.



a) originals singles s1 and s2 respectively

b) mixed signals x1 and x2 respectively.

c) signal output y using only lateral inhibition

d) signal output y using lateral inhibition and masking

Fig. 3 – Signals obtained with real environment.